

**REMARKS**

Reconsideration of this application is respectfully requested.

Claims 35, 37, 39, 41, 43, and 45 were rejected under 35 U.S.C. § 102(e) as allegedly being anticipated by Chang et al. (U.S. Patent No. 6,001,977) and claims 35-46 were rejected under 35 U.S.C. § 103(a) as allegedly being unpatentable over Chang et al. in view of White et al. (U.S. Patent No. 4,677,054). Applicants' claims 35-46 recite methods and kits using probes comprising HIV-1 ORF-1, ORF-4, and ORF-R sequences. The proteins encoded by ORF-1, ORF-4, and ORF-R are now known as the Vpr, Vpu, and Nef proteins of HIV-1. The Examiner alleges that Chang discloses the claimed nucleic acids. The basis for the Examiner's rejection is that the difference between Chang's sequence and applicants' claimed sequence is within a range that can be attributed to sequencing errors, and that the sequences are identical since "the art recognizes that sequencing errors occur in a range between 0.3% and 2.5%, as evidenced by Richterich (Genome Research (1998) 8:251-259)." (Paper No. 32 at 12.) Applicants traverse the rejection.

Richterich (Exhibit 1) does not support the Examiner's position. Richterich does not suggest that the difference between applicants' claimed sequence and Chang's sequence is due to sequencing errors. Rather, Richterich estimates sequencing errors in "raw" DNA sequence data. (Richterich at 251, Title.) As Richterich explains, "raw" DNA sequence data refers to DNA sequences from "large-scale DNA sequencing" projects. (*Id.* at 251, col. 1, ¶ 1.) These are projects where, due to mass-production, many different DNA sequences are generated, but are not subject to any "polishing" by further sequencing efforts. (See *id.* at 251, Abstract.) In the context of these large-

scale DNA sequencing projects, any "resequencing" or "assessments" is considered inefficient and a time delay. (*Id.*) Thus, the error rates of Richterich are based on these mass-production sequencing efforts, not on efforts to sequence a specific DNA clone where polishing, resequencing, and assessments are integral and essential parts of the sequencing effort.

Nonetheless, even in these large-scale DNA sequencing projects, the 1.44 million sequence bases with high quality scores (*i.e.*, good quality sequence) contained only 237 errors, which is a very low error rate. (*See id* at 252, col. 1, ¶ 1.) The Examiner has offered no reasons why error rates of "between 0.3% and 2.5%," as opposed to this substantially lower error rate (approximately 0.017%), should be applicable to Chang's sequences. Applicants note that there is no evidence of record that Chang's sequences are not good quality sequences.

Moreover, neither applicants' sequencing nor Chang's sequencing were the type of large-scale DNA sequencing projects referred to by Richterich, and cannot be considered "raw" DNA sequences. Rather, applicants' and Chang's sequences are "polished" sequences, since "polishing" is an integral and essential part of any sequencing effort. (*See, e.g.*, Current Protocols in Molecular Biology at 7.1.1. (Exhibit 2) and Sambrook et al. at 13.20 (Exhibit 3).)

Chang's sequences would be expected have much lower error rates than the large-scale DNA sequencing projects referred to by Richterich. For example, Chang describes the sequencing of HIV-1 DNA as follows:

Genetic engineering methods are used to determine the nucleotide sequence of HTLV-III DNA. One technique that can be used to determine the sequence is a shotgun/random sequencing method. HTLV-III DNA is sheared randomly into fragments of about 300-500

bp in size. The fragments are cloned, for example, using ml3, and the colonies screened to identify those having an HTLV-III DNA fragment insert. The nucleotide sequence is then generated, with multiple analysis producing overlaps in the sequence. Both strands of the HTLV-III DNA are sequenced to determine orientation. Restriction mapping is used to check the sequencing data generated.

('977 patent at 8, lines 28-39.) To assure a high quality of sequence, Chang indicates that the sequence is "polished" by having multiple analyses producing overlaps and sequencing both strands. Chang's HIV-1 sequences do not contain intact *nef* or *vpr* orfs. ('977 patent at Fig.3.)

Similarly applicants' sequences would be expected have much lower error rates than the large-scale DNA sequencing projects referred to by Richterich. Sequencing of applicants' HIV-1 clone is fully detailed in Wain-Hobson et al. (1995)(Exhibit 4). Wain-Hobson et al. states: "Each nucleotide was sequenced on average 5.3 times: 85% of the sequence was determined on both strands and the remainder was sequenced at least twice from independent clones." (Wain-Hobson et al. at 12, legend to Fig. 1.) Thus, applicants' sequence is a "highly polished" sequence. Applicants' HIV-1 sequence contains intact *nef* and *vpr* orfs. (Specification at 13 and Figs. 9, 11, and 12 and Wain-Hobson et al. at Fig. 1.)

Sequencing the same region multiple times leads to higher accuracy. (Current Protocols in Molecular Biology at 7.1.1.) Also, sequencing both strands leads to higher accuracy. (Sambrook et al. at 13.20.) This additional resequencing was not done in Richterich, but was considered a time delay. Thus, Richterich's error rates of "between 0.3% and 2.5%," are not applicable to a comparison of Chang's and applicants' sequences, which are not "raw" DNA sequences.

Instead, one skilled in the art would have expected that both applicants' and Chang's "polished" sequences would have very low error rates. As Sambrook et al. explains: "When DNA sequencing is carried out carefully, the error rate is less than 0.1%." (Sambrook et al. at 13.20.) There is no reason to believe that applicants' and Chang's sequencing were not performed carefully. Thus, the skilled artisan would have expected error rates of less than 0.1% for applicants' and Chang's sequences. With error rates of less than 0.1%, sequencing errors cannot explain the differences between applicants' and Chang's sequences.

Furthermore, applicants submit herewith Ratner et al. (Exhibit 5) as objective evidence that the differences between applicants' and Chang's sequences are real. Ratner et al. resequenced Chang's clone BH10 (the sequence of which is shown in Figure 3 of the 6,001,977 patent) from both strands. (Ratner et al. at 59, ¶ 4.) Ratner's DNA sequence of BH10 is shown in Figure 1. (*Id.* at 60-61.) Similar to the sequence of BH10 in the '977 patent, Ratner's sequence of BH10 contains a stop codon at position 124 in the 206 codon 3' orf gene (i.e., *nef*). (*Id.* at 61.) Similarly, Ratner's sequence of BH10 contains a frameshift in the *vpr* orf. (*Id.*) Consequently, Chang's BH10 clone does not encode applicants' Nef or Vpr proteins. Ratner et al. indicates that few if any of the sequence differences shown in Figure 1 are likely to represent cloning artifacts or sequencing errors. (*Id.* at 59, ¶ 4.) Consequently, Ratner et al. provides objective evidence that contradicts the Examiner's allegation that the differences between applicants' and Chang's sequences are due to sequencing errors.

Applicants were the first to identify the ORF-R (Nef) and ORF-1 (Vpr) reading frames. Chang was unable to identify these reading frames since Chang's sequences

did not contain the complete *nef* orf because it contained a stop codon and did not contain the complete *vpr* orf because it contained a frameshift. Consequently, applicants' claimed nucleic acids cannot be anticipated by Chang. Accordingly, applicants respectfully request withdrawal of the rejection.

Applicants respectfully submit that this application is now in condition for allowance. In the event that the Examiner disagrees, he is invited to call the undersigned to discuss any outstanding issues remaining in this application in order to expedite prosecution.


Please grant any extensions of time required to enter this response and charge any additional required fees to our deposit account 06-0916.

Respectfully submitted,

FINNEGAN, HENDERSON, FARABOW,  
GARRETT & DUNNER, L.L.P.

Dated: August 1, 2003

By: \_\_\_\_\_

  
Salvatore J. Arrigo  
Registration No. 46,063  
Telephone: 202-408-4160  
Facsimile: 202-408-4400  
E-mail: arrigos@finnegan.com

FINNEGAN  
HENDERSON  
FARABOW  
GARRETT &  
DUNNER LLP

1300 I Street, NW  
Washington, DC 20005  
202.408.4000  
Fax 202.408.4400  
www.finnegan.com

# Estimation of Errors in "Raw" DNA Sequences: A Validation Study

Peter Richterich<sup>1</sup>

Genome Therapeutics Corp., Waltham, Massachusetts 02154 USA

As DNA sequencing is performed more and more in a mass-production-like manner, efficient quality control measures become increasingly important for process control, but so also does the ability to compare different methods and projects. One of the fundamental quality measures in sequencing projects is the position-specific error probability at all bases in each individual sequence. Accurate prediction of base-specific error rates from "raw" sequence data would allow immediate quality control as well as benchmarking different methods and projects while avoiding the inefficiencies and time delays associated with resequencing and assessments after "finishing" a sequence. The program PHRED provides base-specific quality scores that are logarithmically related to error probabilities. This study assessed the accuracy of PHRED's error-rate prediction by analyzing sequencing projects from six different large-scale sequencing laboratories. All projects used four-color fluorescent sequencing, but the sequencing methods used varied widely between the different projects. The results indicate that the error-rate predictions such as those given by PHRED can be highly accurate for a large variety of different sequencing methods as well as over a wide range of sequence quality.

In DNA sequencing, knowledge about the accuracy of sequences can be very valuable. For example, different large-scale sequencing projects may produce sequences at similar rates and costs but with significantly different error rates in the final sequence. One major determinant in the final error rate is the accuracy of the "raw" sequence. Knowledge about the frequency and location of errors in the raw sequence data can help to direct "polishing" efforts to the places where additional effort is needed; it also enables the comparison between different sequencing projects without requiring that the same region be sequenced in each project.

Another area where estimates about sequence error rates would be beneficial is technology development. Accurate error estimates at each base would enable "quality benchmarking" between different methods, thus enabling researchers to choose the method that fills their needs for accuracy and throughput best.

Several groups have developed mathematical models to predict the error probability at any given position in raw sequences. Lawrence and Solovyev used linear discriminant analysis to calculate separate probability estimates for insertions, deletions, and mismatches (Lawrence and Solovyev 1994). Ewing and Green (1998) developed the program

PHRED, which calculates a quality score at each base. This quality score  $q$  is logarithmically linked to the error probability  $p$ :  $q = -10 \times \log_{10}(p)$  (for a discussion of how quality scores are calculated and what the limitations are, see Ewing et al. (1998). When used in combination with sequence assembly and finishing programs that utilize these error estimates, reliable error probabilities promise to increase the accuracy of consensus sequences and to reduce the efforts required in the finishing phase of sequencing projects (Churchill and Waterman 1992; Bonfield and Staden 1995).

To examine the accuracy of probability estimates made by the program PHRED, we compared the actual and predicted error rates for six different cosmid- or BAC-sized projects that were produced by six different large-scale sequencing centers in the United States. All of these six projects used four-color fluorescent sequencing machines; however, the DNA preparation methods, sequencing enzymes, fluorescent dyes and chemistries, and gel lengths varied significantly between the six groups. Table 1 gives an overview of the sequencing projects analyzed. Table 2 lists the different methods used.

## RESULTS

### Error Rate Prediction Accuracy for Six Projects

A comparison of actual and predicted error rates for the six projects in this study is shown in Table 3.

<sup>1</sup>E-MAIL [peter.richterich@genomecorp.com](mailto:peter.richterich@genomecorp.com); FAX (781) 893-9535.

**Table 1. Summary of Data Sets**

Project	Reads	Aligned bases	Average aligned read length
A	455	416,214	915
B	1277	871,230	682
C	1065	603,655	567
D	834	414,595	497
E	1638	1,149,209	702
F	1885	907,796	482
Total	7154	4,362,699	610

The results indicate that PHRED is very successful in identifying bases with low error probabilities. For example, the 1.28 million bases with quality scores of 4–12 (corresponding to error probabilities between 39.8% and 6.3%) contain a total of 187,926 errors. In contrast, the 1.44 million bases with quality scores between 33 and 42 (corresponding to error probabilities between 0.05% and 0.006%) contain only 237 errors, which translates into a 790-fold lower error rate. The trend toward lower error rates can also be observed for each individual project. In most cases, the actual number of errors is close to the predicted error rate. It is also apparent that the actual error rate is typically lower than the predicted error rate.

Both the high overall accuracy and the tendency to slightly overpredict errors are confirmed by statistical analysis, as shown in Table 4. The correlation between predicted and actual error frequencies is excellent for all projects (Spearman correlation coefficient  $>0.89$ ,  $P < 0.0001$ ). Averaged over all projects, the actual error rate is 84.5% of the predicted error rate; the slope of the relation between predicted and actual error rates differs slightly between projects and ranges from 76.6% to 88.4%. To put these differences between projects in relation, it is worthwhile remembering that PHRED quality scores cover a wide dynamic range: The maximum quality score of 51 corresponds to a 50,000-fold lower predicted error rate than the minimum quality score of 4. Even the relative difference between successive quality is larger than the relative difference in the slopes; for example, a quality score of 10 corresponds to an error probability of 10%, whereas a score of 9 corresponds to an error probability of 12.6%.

A different way of looking at the relation between the actual and predicted error rates is shown

in Figure 1. Here, the error rates as a function of the position within all reads in each of the projects, averaged over 50-base windows, is depicted. For all six projects, the predicted error rates are very close to the actual error rates over the entire length of the sequences. Each project has a characteristic distribution of error rates, which differs from each of the other projects. The minimum error rate differs dramatically between projects. The best projects achieve raw error rates of 0.23%–0.36% in the best region of the sequence read, typically from base 150 to 200. The worst project in the data set had an ~10-fold higher error rate of 2.58%.

Toward the end of sequence reads, the error rates increase and start to exceed 10% between bases 300 and 700. In projects that used mainly short gels (e.g., projects D and F), this increase begins sooner, whereas projects that use longer gels show a markedly longer stretch of low error rates (e.g., projects A and B).

Table 5 summarizes key results for the six projects. The first four projects have similar minimum and average error rates. However, the length of the region where the error rate is below 5% differs significantly, from 403 to 682 bases. The project with the shorter low error rate regions contained larger portions of reads generated on short gels, whereas projects A and B were run exclusively on long gels (ABI373 stretch or ABI377 sequencers). Other factors contributing to differences between the first four projects were differences in sequencing chemistries, production scale, and electrophoresis conditions and machines.

Project E and, in particular, project F, had significantly higher error rates than the first four projects. In projects E and F, every sequence generated for the project had been included in the data set, whereas the other four projects had eliminated some "bad" sequences through manual or auto-

**Table 2. Overview of Sequencing Methods Used in the Different Projects**

Template DNA	single-stranded M13, double-stranded plasmids
Sequencing enzymes	Sequenase, <i>Taq</i> , KlenTaqTR, AmpliTaq FS
Sequencing chemistries	Dyes primer (two different dyes chemistries), dye terminator
Sequencing machines	ABI 373, ABI 373 stretch, ABI 377
Gel length	Only short gels, only long gels, mixes of short and long gels

**Table 3. Comparison of Predicted and Actual Error Rates for Six Different Sequencing Projects**

Project	Quality score	4-12	13-22	23-32	33-42	43-51
A	aligned bases	119,246	75,293	70,391	144,876	73,234
	expected errors	20,256	2,064	172	37	1
	actual errors	16,784	1,758	127	17	1
B	aligned bases	182,034	137,940	181,998	399,690	140,176
	expected errors	29,953	3,704	410	102	3
	actual errors	26,038	2,536	287	35	0
C	aligned bases	139,345	131,419	151,197	292,070	68,529
	expected errors	22,277	3,411	357	74	2
	actual errors	16,670	1,513	194	26	3
D	aligned bases	103,898	68,995	68,613	153,730	111,752
	expected errors	16,880	1,919	168	38	3
	actual errors	14,495	1,924	146	59	2
E	aligned bases	378,755	217,438	167,968	392,717	144,313
	expected errors	63,947	6,336	418	95	4
	actual errors	55,968	6,516	355	67	5
F	aligned bases	359,809	136,688	98,840	64,035	5,130
	expected errors	66,938	4,079	256	23	0
	actual errors	57,971	3,856	332	33	1
All	aligned bases	1,283,087	767,773	739,007	1,447,118	543,134
	expected errors	220,252	21,513	1,781	370	13
	actual errors	187,926	18,103	1,441	237	12

matic inspection. After eliminating <10% of the worst sequences in project E, the error rates for the remaining sequences were comparable to those of the first four projects. In contrast, project F showed a much more uniform distribution of sequence quality.

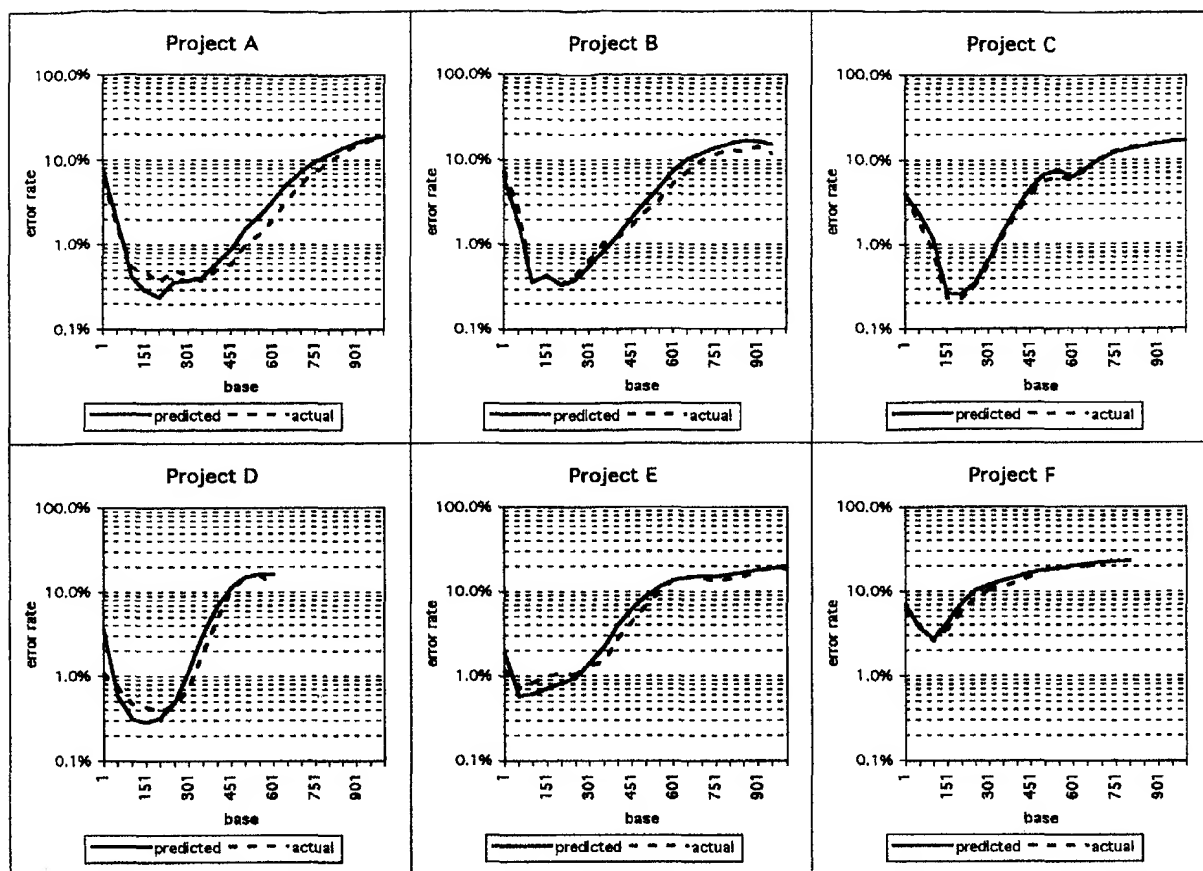
The last column in Table 5 shows the average number of bases with an estimated error probability of at most 0.1%, which is equivalent to a quality score of at least 30. The count of such "very high-quality" bases is a good indicator of sequence quality, both for individual sequences and, when aver-

**Table 4. Summary of Statistical Analysis Results**

Project	Spearman $\rho$	$P >  \rho $	Slope	$t$ ratio	$P >  t $
A	0.9646	<0.0001	0.818	75.1	<0.0001
B	0.9890	<0.0001	0.874	98.2	<0.0001
C	0.9846	<0.0001	0.766	71.6	<0.0001
D <sup>a</sup>	0.8692	<0.0001	0.855	68.3	<0.0001
E	0.9956	<0.0001	0.884	144.3	<0.0001
F	0.9968	<0.0001	0.865	151.6	<0.0001
All	0.9964	<0.0001	0.845	174.5	<0.0001

<sup>a</sup>In project D, the Spearman correlation coefficient  $\rho$  was artificially low as only very few bases (10) bases had a quality score of 5, and none of these bases contained an actual error (expected: 3.16 errors). Exclusion of this quality score gave a Spearman correlation coefficient of 0.9786 ( $P < 0.0001$ ). The frequencies in the slope calculations were weighed by the number of bases at any given quality score and, thus, were not sensitive to such small sample distortions (see Methods).





**Figure 1** Actual and predicted error rates in six different sequencing projects. Actual error rates and predicted error rates in 50-base windows over the length of the sequence reads, averaged over all reads that could be aligned to the consensus sequence by CROSS\_MATCH, are shown. The numbers on the x-axis show the first base in a given 50-base window.

aged over all sequences in a project, as an indicator for the entire project. Compared to the estimated error rates, the count of very high-quality bases is less prone to distortions from a small number of low-quality reads, as the data for project E demonstrate.

#### Prediction Accuracy for Data Subsets of Different Quality

The quality of sequences within any given project can vary substantially, and the use of predicted error rates has the potential to be a powerful tool for qual-

**Table 5. Comparison of Key Results for Six Different Sequencing Projects**

Project	Actual minimum error rate (%)	Actual average error rate (%)	Length of <1% error region	Length of <5% error region	Average bases with $P(\text{error}) < 0.1\%$
A	0.36	3.6	422	682	468
B	0.34	2.8	274	567	395
C	0.23	2.4	291	479	348
D	0.39	3.1	300	403	294
E	0.71	4.7	129	464	317
F	2.58	9.2	0	162	79

ity analysis and control in large-scale DNA sequencing projects. To analyze how accurate PHRED error estimates are for different quality sequences within the same sequencing project, we subdivided a data set into four quartiles, based on the number of very high-quality bases in each sequence (see Methods). The comparison of actual and predicted error rates is shown in Figure 2.

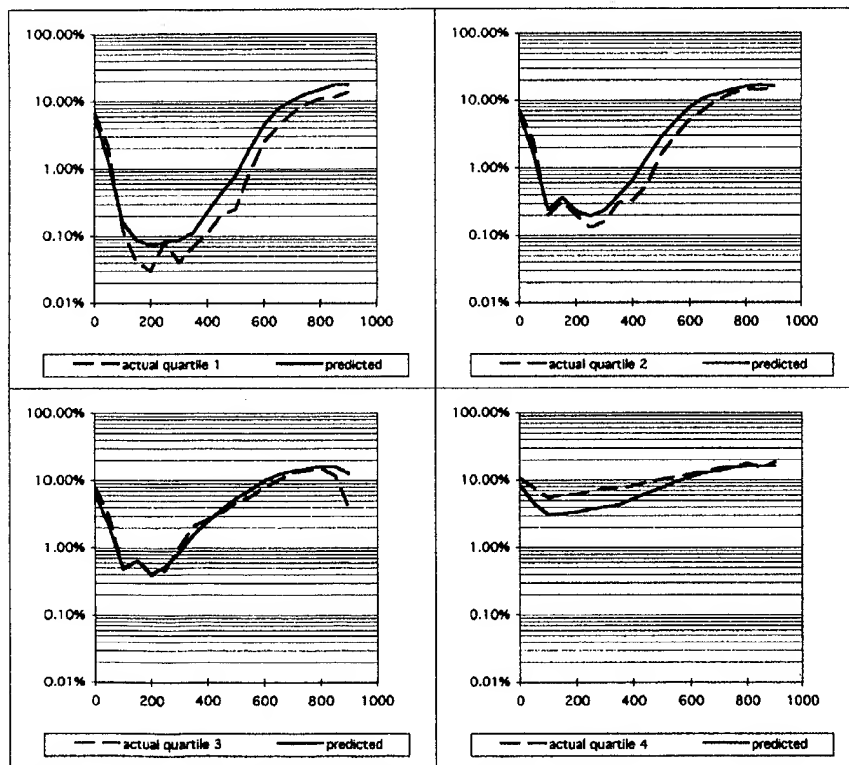
When measured by the error rate in the best region of a sequence, the data quality in the different quartiles varies >100-fold between the best and the worst 25% of the sequences. The best quartile showed ~0.03% error for >100 bases, whereas the error rate in the worst quartile always exceeded 5%. In quartiles 2 and 3, the predicted error rates match the actual error rates very closely. In the best and

worst quartiles, PHRED's accuracy was somewhat lower from base 100 to 500. In the best sequences, PHRED's error estimates were about twofold too high; in the worst sequences, the error estimates were too low, again by a factor of 2. This underprediction of errors can be partially explained by the fact that PHRED gives ambiguous base calls (N's) a quality score of 4, corresponding to an error probability of 39.8%; however, N's will always show up as an actual error. Even in the worst and best quartiles, however, the predicted error rate curves are very similar to the actual error rate curves.

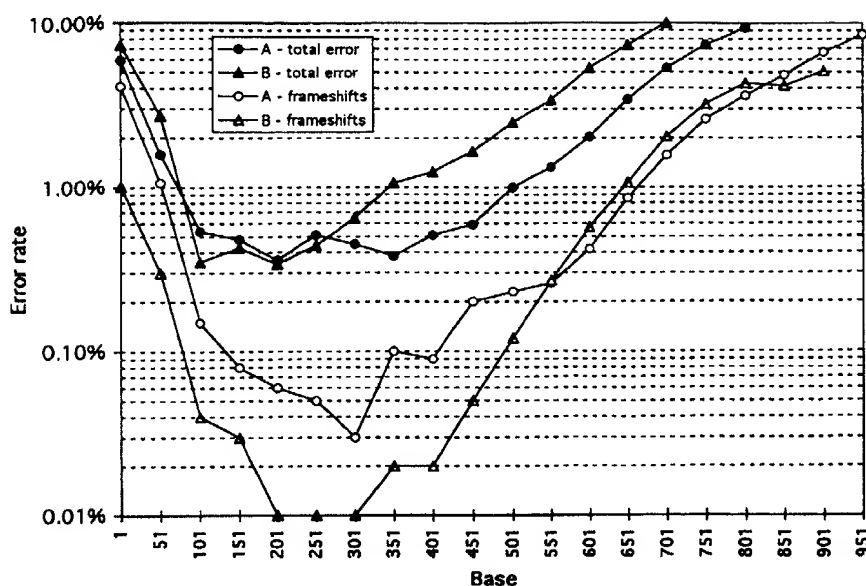
The results shown in Figure 2 also demonstrate that the count of very high-quality bases, or bases with an estimated error probability of at most 0.1%, can be used effectively to characterize the overall

quality of a sequence read. Sorting the sequence reads into quartiles based on the number of very high-quality bases worked well, as shown by the >100-fold difference in the minimum error rate between the first and the fourth quartile.

Other methods to characterize the overall quality of individual reads based on PHRED quality scores can give similar results. For example, counting bases above a minimum quality threshold anywhere in the range of 20–40 gave similar results for most data sets (not shown), and such counts are used by a number of different laboratories as quality measures. Alternatively, the quality values can be converted to error probabilities and averaged to give the predicted error rate for the trace, or summed to give the total predicted number of errors in a trace. However, such averages and totals can sometimes give a misleading picture, as the following example illustrates. Assume that two sequence reads have very similar quality in the alignable part of the read but that one of the two sequences was run much longer and



**Figure 2** Actual and predicted error rates in different quality subsets of project B. Sequence reads were sorted by the number of bases with a predicted error rate of at most 0.1% (very high-quality bases), and assigned to quartiles, with quartile 1 corresponding to the highest numbers. Actual and predicted error rates for all sequences in each subset were calculated as in Fig. 1. Note that a number of sequence reads that had been rejected because of too low quality were added back to the data set for illustrative purposes, all of which are in quartile 4. These sequences were not included in the data sets used to generate Figs. 1 and 3 and Tables 1 and 3.



**Figure 3** Actual frameshift and total error rates for projects A and B. To calculate frameshift error rates, only insertions and deletions were counted. Mismatch errors, which account for the vast majority of errors after base 150, were included only in the total error count. Note that project B ( $\blacktriangle, \triangle$ ) has a slightly similar or slightly higher total error rate compared to project A ( $\bullet, \circ$ ) but only about one-third as many insertions and deletions up to base 500. For both projects, the frameshift error rate in the raw data is  $<1$  in 1000 for  $>300$  bases, and  $\leq 1$  in 10,000 for  $>100$  bases in project B.

therefore contains a longer unalignable "tail" of very low-quality bases. When calculating the average error rate for these two sequences, the second sequence will have a much higher average error and, therefore, appear to be of lower quality. In contrast, the counts of very high-quality bases for both sequences will be very similar, as the unalignable tails contain few, if any, high-quality bases. Therefore, counts of bases above a high enough quality threshold will give a more robust and clearer picture of trace quality.

#### Frameshift Error Rates for Different Sequencing Chemistries

Depending on how biologists use DNA sequences, knowledge about total error rates in raw sequences may or may not be sufficient. For example, frameshift errors in coding sequences will generally lead to incorrectly predicted open reading frame, whereas mismatch errors will do so only if the mismatch introduces a stop codon or a new splice site. At the time of this writing, PHRED did not differentiate between mismatch and frameshift errors, but only estimated total error rates. This might occa-

sionally lead to questionable conclusions, as the results shown in Figure 3 illustrate.

Figure 3 shows the total actual error rates and the frameshift error rates for two projects, A and B. The total error rates for both projects are similar for up to 350 bases; after 350 bases, project B has a somewhat higher total error rate. However, examining the frameshift error rate gives rise to a different picture: from base 1 to 500, project A has approximately four times as many insertions and deletions as project B. This difference in frameshift error rates can be explained by the sequencing chemistries that were used in the two projects. Project B, with the lower frameshift error rate, used only dye terminator chemistry, which is known to eliminate band spacing artifacts from hairpin structures ("compressions"). Project A, on the

other hand, used dye primer chemistry, which is more prone to insertion and deletion errors from mobility artifacts, for most sequencing reactions.

#### DISCUSSION

As large-scale DNA sequencing has become a more routine and common process, the traditional methods for assessing sequence quality have become unsatisfactory. In projects like single-pass cDNA sequencing, it is not possible to calculate and compare error rates after finishing a sequence, as finishing never takes place. Even when a comparison between raw and finished sequence can be done, the time delay between raw data generation and quality assessment is often large. This delay makes it difficult to improve ongoing projects, and it sometimes makes it impossible to capture problems early on. Some immediate quality feedback can be reached by including known standard sequences for quality control. However, this approach can be costly, and it fails when error profiles differ between standard and unknown sequences.

In contrast to these traditional methods to assess sequence accuracy, direct estimation of error

rates in raw sequence data would enable immediate quality control and feedback. Accurate, base-by-base estimates of error probabilities could also increase the utility of single-pass sequences significantly, allow efficient comparison and optimization of different sequence chemistries, and enable the development of better software tools for sequence assembly and analysis.

The critical question for any error rate prediction tool is how accurate are the error rate estimates, in particular if different sequencing methods and chemistries are used? The results presented herein provide an answer to this question for the program PHRED, as well as clues where further development would be useful. As shown in Tables 3 and 4 and in Figure 1, the agreement between predicted and actual error rates was very good in each of the six different projects analyzed. The observed high level of prediction accuracy in all of these projects is almost astonishing if one takes into account that actual errors are binary (a base is either correct or wrong), whereas predicted error rates are probabilities on a scale from 0.0 to 1.0. The observed tendency to overpredict error rates can be at least partially explained by the "small sample correction" that was used in the derivation of threshold parameters for quality scores (Ewing and Green 1998). For most practical applications, such a somewhat conservative estimation of quality scores is tolerable or even desirable. Overall, the results clearly show that error probabilities given by PHRED accurately describe raw sequence data quality.

In judging the usefulness of predicted error probabilities, it is important to know how differences in sequencing methods will influence the prediction accuracy. For example, the larger variation in peak heights tends to be larger in dye terminator sequencing than in dye primer sequencing, and different sequencing enzymes are known to produce different specific height variation patterns. Any estimation of error probabilities that takes the peculiarities of a specific sequencing chemistry into account would therefore be expected to be less accurate for different chemistries.

The projects included in this study were specifically chosen to provide an initial answer to the question of how generally useful PHRED quality scores are. These projects represent the vast majority of different multicolor fluorescent sequencing methods used in the last 3 years: different template DNAs and DNA preparation methods, different enzymes, gel lengths, run conditions, and different fluorescent dyes. The data also include a considerable spread in data quality, both between projects

and within individual projects. None of the projects analyzed here were included in PHRED's training set, and just one of the six laboratories that contributed data to this study also contributed data to the training data sets. One of the projects in this study consisted entirely of dye terminator sequences, which presented only a small fraction of the sequences in the test data set. Another project exclusively used a set of fluorescent dyes different from those used in the training sets. Each project differed from the other projects in this study in at least one, and typically many, experimental aspects like template preparation, sequencing enzymes, gel run conditions, and so forth. Despite these differences, the accuracy of error rate predictions was very similar for all projects.

Our results justify some optimism about the accuracy of PHRED quality scores for minor changes in sequencing technology, for example, sequences generated by new enzymes and fluorescent dyes. Initial studies showed that PHRED quality scores were also accurate for sequences produced by multiplex sequencing with radioactive detection (P. Richterich, unpubl.). However, we also observed two effects that can invalidate PHRED quality scores during these studies. First, sequences generated by chemical sequencing gave too low quality scores at mixed (A + G) reactions. Because secondary peak height is one of the parameters used in the error rate predictions, this is not surprising. Another potential source of error is high-frequency noise in the trace data. With such data, PHRED occasionally underestimated the band spacing by a factor of 2 or more, which resulted in incorrect base calls and quality scores. By applying simple smoothing algorithms to data with high-frequency noise, these problems could typically be resolved. Similar steps may be necessary to obtain accurate PHRED quality scores on data that have been generated by different sequencing instruments or preprocessed by different software.

Accurate quality scores can have a major impact on how sequences are used downstream from the sequence production process. In traditional sequencing projects where the goal is complete coverage at a final error rate below (e.g.) 1 in 10,000, the accuracy goals can be reached with single sequence reads as long as the quality scores are at least 40 (however, other potential problems like clone instability may make higher coverage advisable). Interesting questions arise as to how individual read quality contributes to project quality, or the error rate of the "final" sequence. Under the assumption that errors between different sequence reads are

completely independent, one could argue that two reads with a quality score of 20 (error probability of 1 in 100) are just as valuable as one sequence with a quality score of 40 (error probability of 1 in 10,000). However, although a single sequence stretch with quality levels above 40 would give a final sequence with an error rate of  $<1$  in 10,000, assembling a consensus from two sequences with quality scores of 20 (1% error rate) could lead to one of two results: If the errors were completely random, the consensus sequence would be ambiguous at 2% of all locations; if the errors were completely localized, for example, because of reproducible compressions, the consensus sequence would have one "hidden" error every 100 bases. Typically, consensus sequences derived from low-quality sequences will have both kinds of problematic regions. Increased coverage can rapidly eliminate the random errors; however, increased coverage does not resolve errors from systematic sources. Manual examination of such problem areas is generally required; such "contig editing," however, tends to be time consuming, requires highly trained personnel, is an obstacle toward complete automation of DNA sequencing, and sometimes fails to eliminate all errors. This leads to the somewhat counterintuitive conclusion that the practical value of increasing sequence quality can be even higher than indicated by the quality scores: One sequence of average quality above 40 can be "worth" more than two sequences of average quality 20.

Another application of DNA sequencing where high quality can be of disproportionately high value is the search for mutations in genomic DNA. In low quality sequences, secondary peaks and low resolution often complicate the identification of heterozygous mutations. In regions of higher sequence quality, such secondary peaks are smaller or absent and peaks are better resolved. Therefore, both false-positive and false-negative errors can be significantly reduced in high-quality regions. Tools like PHRED, which can accurately measure sequence quality from trace data, can be of twofold value for mutation detection. First, base-specific quality scores can allow optimization of sequencing methods and strategies for mutation detection. Second, the quality scores can be used to evaluate the usefulness of individual sequence reads for mutation detection (e.g., by discarding reads below minimum thresholds), and they can guide software that automatically detects mutations.

The ability to predict error rates in a highly accurate fashion is likely to have a major impact in applications like those described above. PHRED is

the first widely used program that accurately predicts base-specific error probabilities. However, the algorithm for determining quality values has been described (Ewing and Green 1998), and it should be straightforward to implement similar quality values in other base-calling programs. Furthermore, an extension of the approach developed by Ewing and Green should be possible. For example, differentiation between mismatch and frameshift errors would enable better comparisons of sequencing methods with similar total error rates but different frameshift error rates. Several groups have described efforts to calculate separate probabilities (or "confidence assessments") for mismatch errors and frameshift errors (Lawrence and Solovyev 1994; Berno 1996). Their results demonstrated that different approaches to error type characterization are feasible and promising. Implementation of such error type predictions in other programs similar to the way PHRED uses quality scores would enable better method assessments, benchmarking, and production quality control, and could have a significant impact on downstream uses of DNA sequence information.

## METHODS

### Data Sets

For one project, sequence raw data in the form of ABI trace files were downloaded from a public FTP site. Sequence data for the five other projects were kindly provided by five different large-scale sequencing groups. Table 1 gives a summary of the six projects, and Table 2 gives an overview of the different sequencing methods used in the projects. The projects differed in the amount of prescreening of data that had been done, reflecting different approaches to quality control in different laboratories. In two projects (B and C), different software programs had been used to identify and eliminate low-quality sequences. One project (F) included all data files generated, whereas the other three projects had excluded "failed lanes."

### Comparison of Actual and Predicted Error Rates

The sequences for all traces in each project were recalled using the program PHRED (v. 961028). Next, sequences in each project were assembled with PHRAP (P. Green, unpubl.). Slightly different methods were chosen for the statistical and graphical evaluation of the error rate prediction accuracy. In the statistical evaluation, only the longest contig produced by PHRAP was considered. The tables of aligned bases and observed discrepancy counts for

each quality score were taken from the PHRAP output and analyzed as follows. The expected number of discrepancies ( $E$ ) at each quality score ( $q$ ) was calculated by multiplying the number of aligned bases ( $N$ ) with the error probability corresponding to the quality score:  $E = N 10^{-0.1q}$ . The Spearman ranking coefficients were calculated by comparing the expected and observed error frequencies. To obtain the quantitative relation between the expected and observed error rates over the entire range, a least-squares fit between the observed and expected rates was performed, with the intercept set to zero and the number of aligned bases at each quality score used as weights.

For a graphical comparison of estimated and actual error rates in 50-bp windows, the following steps were taken. For two of the projects, the consensus sequence was retrieved from public databases. For the four other projects, the DNA sequence and quality information were used by the program PHRAP to assemble consensus sequences for each of the projects. The individual reads were aligned to the consensus sequences of the longest contig, using the program CROSS\_MATCH (P. Green, unpubl.), after removing single-coverage regions from the ends of the consensus sequence. CROSS\_MATCH uses an implementation of the Smith-Waterman algorithm to generate alignments that typically do not include the ends of sequences, where disagreements are commonly due to vector sequence or low quality sequence.

The quality files generated by PHRED and the alignment summaries generated by CROSS\_MATCH were then analyzed as follows. First, the region of each query sequence that had been aligned by CROSS\_MATCH was determined. Next, the actual and predicted error rates for the entire aligned part of each individual sequence was calculated. In addition, the average actual and predicted error rates for all alignable sequences together were calculated for windows of 50 bases in length. To calculate the predicted error rate, the quality scores  $q$  determined by PHRED at each base were converted to error probabilities as described above (Ewing and Green 1998).

#### Subdividing Data into Subsets Based on Data Quality

To examine the accuracy of PHRED quality scores for data subsets of different quality within a project, the following approach was taken. For all sequence reads in project B, the number of bases with a quality score of at least 30 in each sequence was determined (bases with quality scores of at least 30 were called very high-quality bases, or VHQ bases). Se-

quences were sorted in descending order based on the number of very high-quality bases, and divided into four quartiles. Accordingly, quartile 1 contained 25% of sequences with the highest number of very high-quality bases, and quartile 4 contained the "worst" sequences. To illustrate the prediction accuracy in data with relatively high error rates, sequences from project B that had been "discarded" because they had not met the minimum quality criteria were added back to the data set. The sequences in each quartile were compared to the consensus sequences that had been generated using the entire data set, as described above for the graphical comparison.

#### Determining Actual Frameshift Error Rates

The calculation of actual frameshift error rates in the raw sequence data was performed using CROSS\_MATCH, similar to the procedure described above for total error rates, except that only insertion and deletion errors were counted. Because PHRED does not give separate frameshift error estimates, a comparison of predicted and actual frameshift errors is not possible.

#### ACKNOWLEDGMENTS

I thank the participating laboratories for contributing their data, Dr. Josée Dupuis for help with the statistical analysis, and Dr. Phil Green for helpful discussions.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

#### REFERENCES

- Berno, A.J. 1996. A graph theoretic approach to the analysis of DNA sequencing data. *Genome Res.* **6**: 80-91.
- Bonfield, J.K. and R. Staden. 1995. The application of numerical estimates of base calling accuracy to DNA sequencing projects. *Nucleic Acids Res.* **23**: 1406-1410.
- Churchill, G. and M.S. Waterman. 1992. The accuracy of DNA sequences: estimating sequence quality. *Genomics* **14**: 89-98.
- Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using *phred*. II. Error probabilities. *Genome Res.* (this issue).
- Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment. *Genome Res.* (this issue).
- Lawrence, C.B. and V.V. Solov'yev. 1994. Assignment of position-specific error probability to primary sequence data. *Nucleic Acids Res.* **22**: 1272-1280.

Received October 27, 1997; accepted in revised form February 3, 1998.

# DNA Sequencing Strategies

## UNIT 7.1

This unit contains a general discussion of factors that should be considered before embarking on a DNA sequencing project. In general, any sequencing strategy should include plans for sequencing both strands of the DNA fragment. Complementary strand confirmation leads to higher accuracy, especially when sequencing regions where artifacts such as "compressions" are a problem. Sequencing the opposite strand is often required to obtain accurate data for such regions.

The most commonly used methods for generating appropriately sized DNA fragments for dideoxy and chemical sequencing are discussed below. The biochemistry underlying these procedures, as well as how to choose between these and alternative sequencing methods, are discussed in the introduction to this chapter.

### DIDEOXY SEQUENCING

#### Planning for Dideoxy Sequencing

Sequencing determination of a fragment of <500 nucleotides is usually straightforward because this amount of sequence information can reliably be determined from a single set of sequencing reactions. Fragments of this size can usually be subcloned directly into an appropriate single- or double-stranded DNA sequencing vector. If the vector generates single-stranded DNA, such as the M13mp vectors described below, the fragment should be cloned in both orientations so that both strands of the insert are produced as single-stranded DNA. A primer that hybridizes to a site on the vector adjacent to the insert DNA is then used to sequence both clones, generating the sequence of each strand. When sequencing double-stranded plasmid DNA, there are two options for obtaining the sequence of each strand. A single primer can be used if the insert DNA is cloned in both orientations. Alternatively, two primers that hybridize to plasmid sequences on opposite sides (and opposite strands) of the insert DNA can be used to sequence a single clone.

To sequence larger regions of DNA completely, it is generally necessary to subdivide a large fragment into smaller ones that can then be individually sequenced. Three general approaches are currently used. In the first approach, known as "shotgun cloning," random

fragments are created from longer DNA fragments by physical shear, digestion by nucleases, (e.g., DNase I) or by restriction digests with endonucleases that make frequent cuts in the fragment (e.g., those with four-base recognition specificity; Frischauf et al., 1980; Anderson, 1981; Bankier and Barrell, 1983; Bankier et al., 1988; Hong, 1982; Messing, 1983, 1988; Deininger, 1983a, 1983b; Bankier, 1984; Lin et al., 1985). These fragments are combined and the entire pool is ligated into the appropriate sequencing vector. After the DNA sequence of the various cloned fragments has been determined, the final sequence is compiled by computer using overlapping information from the individual fragments (UNIT 7.7). However, with more complex (i.e., longer) sequences, this approach becomes tedious since it requires purifying, ligating, and cloning numerous individual fragments. In addition, finding the final few percent of a sequence by this procedure can consume a disproportionately large amount of time.

A second subcloning strategy for sequencing large DNA fragments is to generate an ordered set of subclones from a large DNA molecule. This is usually done by making progressive (nested) sets of deletions from a clone containing the entire DNA fragment to be sequenced. Numerous protocols exist for making nested deletions by enzymatic means; two such protocols using exonuclease III (Henikoff, 1984; Guo and Wu, 1982; Okita, 1985; Ozkaynak and Putney, 1987; Smith, 1979, 1980; Strauss and Zagurski, 1991) and nuclease *Bal* 31 (Guo and Wu, 1982; Guo et al., 1983; Vocke and Bastia, 1983; Yanisch-Perron et al., 1985; Poncz et al., 1982; Misra, 1985) are presented in UNIT 7.2. Another enzymatic method for making nested deletions utilizes T4 DNA polymerase (Dale et al., 1985). Other methods for isolating nested sets of deletion fragments include size-selection of inserts (Barnes, 1987; Barnes and Bevan, 1983; Barnes et al., 1983; Vocke and Bastia, 1983), oriented restriction fragment subcloning (Yanisch-Perron et al., 1985; Lee and Lee, 1989; Hoheisel and Pohl, 1986), transposon-mediated deletions (Ahmed, 1987a, 1987b; Nag et al., 1988; Peng and Wu, 1986), progressive synthesis (Burton et al., 1988; Liu and Hackett, 1989), and oligonucleotide-directed mutagenesis (Shen and Wayne, 1988; Wang et al., 1989). Commercial kits for

### DNA Sequencing

#### 7.1.1

Contributed by Barton E. Slatko, Richard L. Eckert, Lisa M. Albright, and Frederick M. Ausubel

*Current Protocols in Molecular Biology* (1999) 7.1.1-7.1.7

Copyright © 1999 by John Wiley & Sons, Inc.

Supplement 46

**2**

# ***Molecular Cloning***

**A LABORATORY MANUAL**  

---

**SECOND EDITION**

---

**J. Sambrook**

UNIVERSITY OF TEXAS SOUTHWESTERN MEDICAL CENTER

**E.F. Fritsch**

GENETICS INSTITUTE

**T. Maniatis**

HARVARD UNIVERSITY



**Cold Spring Harbor Laboratory Press  
1989**



## **Molecular Cloning**

A LABORATORY MANUAL  
SECOND EDITION

All rights reserved  
© 1989 by Cold Spring Harbor Laboratory Press  
Printed in the United States of America

9

*Book and cover design by Emily Harste*

*Cover:* The electron micrograph of bacteriophage  $\lambda$  particles stained with uranyl acetate was digitized and assigned false color by computer. (Thomas R. Broker, Louise T. Chow, and James I. Garrels)

*Cataloging in Publications data*

Sambrook, Joseph

Molecular cloning : a laboratory manual / E.F.

Fritsch, T. Maniatis—2nd ed.

p. cm.

Bibliography: p.

Includes index.

ISBN 0-87969-309-6

1. Molecular cloning—Laboratory manuals. 2. Eukaryotic cells—Laboratory manuals. I. Fritsch, Edward F. II. Maniatis, Thomas III. Title.

QH442.2.M26 1987

574.87'3224—dc19

87-35464

Researchers using the procedures of this manual do so at their own risk. Cold Spring Harbor Laboratory makes no representations or warranties with respect to the material set forth in this manual and has no liability in connection with the use of these materials.

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Cold Spring Harbor Laboratory Press for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that the base fee of \$0.10 per page is paid directly to CCC, 21 Congress St., Salem MA 01970. [0-87969-309-6/89 \$00 + \$0.10] This consent does not extend to other kinds of copying, such as copying for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale.

All Cold Spring Harbor Laboratory Press publications may be ordered directly from Cold Spring Harbor Laboratory Press, 10 Skyline Drive, Plainview, New York 11803. Phone: 1-800-843-4388. In New York (516) 367-8423. FAX: (516) 367-8432.

to search sequences for regions that are complementary to synthetic oligonucleotides.

#### *Accuracy of the sequence*

When DNA sequencing is carried out carefully, the error rate is less than 0.1%. However, to achieve this high rate of accuracy, it is necessary to sequence both strands of the target DNA completely and to resolve all ambiguities and discrepancies. In this respect, random sequencing has an advantage, since the gradual accumulation of redundant primary sequences greatly improves the accuracy of the final assembled sequence. However, there may be regions of the target DNA that cannot be sequenced accurately by either the random method or directed methods. Resolving these difficult sequences often takes a surprisingly long time and sometimes requires the use of base analogs (to eliminate compressions) or Maxam-Gilbert sequencing.

#### *Future direction of the project*

Different sequencing strategies yield different types of material that can be used in later experiments. For example, nested sets of deletions generated for DNA sequencing can be used to study the domains within a promoter region or sets of oligonucleotides complementary to different regions of the target fragment can be used to sequence mutant forms of the target sequence. Random subclones created for shotgun sequencing provide a store of material that can subsequently be used for site-directed mutagenesis or for the generation of radiolabeled probes.

# Nucleotide Sequence of the AIDS Virus, LAV

Simon Wain-Hobson,\* Pierre Sonigo,\*

Olivier Danos,† Stewart Cole,‡ and Marc Alizon§

\* Unité de Recombinaison et Expression Génétique

† Unité des Virus Oncogènes

‡ Groupement de Génie Génétique

§ Unité d'Oncologie Virale

Institut Pasteur

25 et 28 rue du Dr. Roux

75724 Paris, Cedex 15, France

## Summary

The complete 9193-nucleotide sequence of the probable causative agent of AIDS, lymphadenopathy-associated virus (LAV), has been determined. The deduced genetic structure is unique: it shows, in addition to the retroviral gag, pol, and env genes, two novel open reading frames we call Q and F. Remarkably, Q is located between pol and env and F is half-encoded by the U3 element of the LTR. These data place LAV apart from the previously characterized family of human T cell leukemia/lymphoma viruses.

## Introduction

The recent onset of severe opportunistic infections among previously healthy male homosexuals has led to the characterization of the acquired immune deficiency syndrome (AIDS) (Gottlieb et al., 1981; Masur et al., 1981). The disease has spread dramatically, and new high-risk groups have been identified: patients receiving blood products, intravenous drug addicts, and individuals originating from Haiti and Central Africa (Piot et al., 1984). AIDS is a fatal disease, and there is at present no specific treatment. The causative agent was suspected to be of viral origin since the epidemiological pattern of AIDS was consistent with a transmissible disease, and cases had been reported after treatment involving ultrafiltered anti-hemophilia preparations (Daly and Scott, 1983). A decisive step in AIDS research was the discovery of a novel human retrovirus called lymphadenopathy-associated virus (LAV) (Barré-Sinoussi et al., 1983). The properties of the virus consistent with its etiological role in AIDS are: the recovery of many independent isolates from patients with AIDS or related diseases (Montagnier et al., 1984); high LAV seropositivity among these populations (Brun-Vézinet et al., 1984); a tropism and cytopathic effect in vitro for the helper/inducer T-lymphocyte subset T4 (Klatzmann et al., 1984), also found depleted in vivo.

Other groups have reported the isolation of human retroviruses, the human T cell leukemia/lymphoma/lymphotropic virus type III (HTLV-III) (Popovic et al., 1984) and the AIDS-associated retrovirus (ARV), which display biological and sero-epidemiological properties very similar to if not identical with those of LAV (Levy et al., 1984; Popovic et al., 1984; Schüpbach et al., 1984). Both LAV and HTLV-

III genomes have been molecularly cloned (Alizon et al., 1984; Hahn et al., 1984). Their restriction maps show remarkable agreement, including a Hind III restriction site polymorphism, bearing in mind the variability of this virus (Shaw et al., 1984) and confirming that these two viruses represent a single viral lineage.

In addition to its obvious diagnostic and therapeutic potential, the LAV DNA nucleotide sequence is essential to an understanding of the genetics and molecular biology of the virus and its classification among retroviruses. We report here the complete 9193-nucleotide sequence of the LAV genome established from cloned proviral DNA.

## Results

### DNA Sequence and Organization of the LAV Genome

We have reported previously the molecular cloning of both cDNA and integrated proviral forms of LAV (Alizon et al., 1984). The recombinant phage clones were isolated from a genomic library of LAV-infected human T-lymphocyte DNA partially digested by Hind III. The insert of recombinant phage  $\lambda$ J19 was generated by Hind III cleavage within the R element of the long terminal repeat (LTR). Thus each extremity of the insert contains one part of the LTR. We have eliminated the possibility of clustered Hind III sites within R by sequencing part of an LAV cDNA clone, pLAV 75 (Alizon et al., 1984), corresponding to this region (data not shown). Thus the total sequence information of the LAV genome can be derived from the  $\lambda$ J19 clone.

Using the M13 shotgun cloning and dideoxy chain termination method (Sanger et al., 1977), we have determined the nucleotide sequence of  $\lambda$ J19 insert. The reconstructed viral genome with two copies of the R sequence is 9193 nucleotides long. The numbering system starts at the cap site (see below) of virion RNA (Figure 1).

The viral (+) strand contains the statutory retroviral genes encoding the core structural proteins (gag), reverse transcriptase (pol), and envelope protein (env), and two extra open reading frames (orf) that we call Q and F (Table 1). The genetic organization of LAV, 5'LTR-gag-pol-Q-env-F-3'LTR, is unique. Whereas in all replication-competent retroviruses pol and env genes overlap, in LAV they are separated by orf Q (192 amino acids) followed by four small (<100 triplets) orf. The orf F (206 amino acids) slightly overlaps the 3' end of env and is remarkable in that it is half-encoded by the U3 region of the LTR.

Such a structure clearly places LAV apart from previously sequenced retroviruses (Figure 2). The (-) strand is apparently noncoding. The additional Hind III site of the LAV clone  $\lambda$ J81 (with respect to  $\lambda$ J19) maps to the apparently noncoding region between Q and env (positions 5166-5745). Starting at position 5501 is a sequence (AAGCCT) that differs by a single base (underlined) from the Hind III recognition sequence. It is anticipated that many of the restriction site polymorphisms between different isolates will map to this region.

Kim H.

100

516

100

**CONCLUSIONS**

348

612

GTG

**ITCA**

901

111

11 15  
21 15

**LED**

2014

**0400**

• • •

**LAGA**

AGAC

**TCAT**

**GAG**

CCG

CTAD

LeuGlyIlelleGlnAlaGlnProAspLysSerGluSerGluLeuValAsnGlaullelleGluGlnLeulleLysLysGlyLysValTyrLeuAlaTrpValProAlaHisGlyCylle  
ATTAGCAATATTTCACGCCAACCGATATAAAGTGAACTACAGCTTAGTCAAATAAATACAGCACTTAATAAAAAAGCAAAGGTCTATCTGGCATGGCTACAGCACCAAGGAAT  
3700

GlyGlyAsnGlnGlnValAspLysLeuValSerAlaGlyIleArgLysValLeuPheLeuAspGlyLysLeuAlaGlnAspGluHisGlyLysTyrHisSerAsnTrpArgAlaMet  
TGCAGCAATCAACAAGTATAGAATAATAGCTAGCTGCTGCAATACGAAAGTACTATTTTTATAGTGAATAGATAAGCCCAAGTACGAACATGCAAAATACAGTAATTCGAGACCAAT  
3800

AlaSerAspPheAsnLeuProProValValAlaLysGluIleValAlaSerCysAspLysCysGlnAlaMetHisGlyGlnValAspCysSerProGlyLysTrpGln  
GGCTAGTGAATTTTAACTGCCACCTGTAGTACCAAAAGAAATAGTACCACGCTGTGATAAATGTCAGCTAAAAGCAGACGCCATGCTATGGCAAGTAGTACTGTAGTCCAGGAATATGCCA  
3900

LeuAspCysThrHisLeuGluGlyLysValIleLeuValAlaValHisValAlaSerGlyTrpLysIleGluAlaGluValIleProAlaGluThrGlyGlnGluThrAlaTyrPheLeuLeu  
ACTAGATTGTACACATTAGCAAGGAAAGCTTATCTCTGCTAGCAGCTTCATGTAGCAGCTGCATATATACAGCAGCAAGCTTATTTCAGCAGCAAGCAGGCCGAAACAGCATCTATCTCTTT  
4000

LysLeuAlaGlyArgTrpProValLysThrIleHisThrAspAsnGlySerAsnPheThrSerThrThrValLysAlaAlaCysTrpTrpAlaGlyLysLysGlnGluPheGlyIlePro  
AAAATTAGCAGCAAGTATGCCATGAAACAAATACATACAGCAATGCCAGCAATTTCCAGCATCTACGCTTAAGCCGCCCTGTTGCTGGCGCGCAATCAACAGCAATTTGCAATTCG  
4100

TyrAsnProGlnSerGlnGlyValValGluSerMetAsnLysGluLeuLysIlelleGlnAlaArgAspGlnAlaGlyLysGlyLysThrAlaValGlnMetAlaValPheIle  
CTCAATCCCCAAAGTCAAGGAGTGTAGAACTATGAAATAAGAAATTAAGAAAAATATAGCCGAGTACAGCATGAGCGCTGAACATCTTAAGCAGCAGTCAAGTAATGCCATATTCAT  
4200

HisAsnPheLysArgLysGlyGlyIleGlyGlyTyrSerAlaGlyGluArgIleValAspIleAlaThrAspIleGlnThrLysGluLeuGlnLysGlnIleThrLysIleGlnAsn  
CCCAATTTTAAACAGAAAGCGGGGATTGCGGGCTACAGTGCAGCGGCAAGAACTAGACATAAATACCAACAGACATACAACTAAAGCAATACAAAAACAAATACAAAAATTTCAAAA  
4300

PheArgValTyrTyrArgAspSerArgAspProLeuTrpLysGlyProAlaLysLeuLeuTrpLysGlyGluGlyAlaValValIleGlnAspAsnSerAspIleLysValValEroArg  
TTTTGGGCTTTATACAGCCGACGACAGATTCACCTTGGCAAGGACGCAAGAGCTCTCTGCAAGAGTGAAGCCGAGTAGTAATACACATAATAGTCACATAAAAGTACGCGAAC  
ORF Q = CysGlu  
4500

ArgLysAlaLysIlelleArgAspTrpGlyLysGlnMetAlaGlyAspAspCysValAlaSerArgGlnAspGluAsp  
GlyLysGlnArgSerLeuGlyLysIlelleGlnAlaGlyGlyGlnGlyGlnValMetLysIleGlnAlaArgArgMetArgIleArgThrTrpLysSerLeuValLysHisHisMetTyrValSer  
ACAAAGCAACAGCATCTATGGCATATTGCAAAACAGTTCGCAGGTGTGATGTTCTGTCGCACTAGACAGCATGACGATTAGCAACATGCAAAAGTTTACTTAAACAGCATATGATGATTAT  
4600

GlyLysAlaArgGlyTrpPheTyrArgHisHisTyrGJserProHisProArgIleSerSerGluValHisIleProLeuGlyAspAlaArgLeuValIleThrThrTrpTrpGlyLeu  
CAGCGAAAGCTAGCGGATGCTTTTATAGACATCCTATGAAAGCCCTCATCCAAGATAAGTTCAGAACTACACATCCCATCTAGCGGATGCTAGATTGCTAATAACACATATTCGGGCT  
4700

HisThrGlyGluArgAspTrpHisLeuGlyGlnGlyValSerIleGluTrpArgLysLysArgTyrSerThrGlnValAspProGluLeuAlaaspGlnLeuIleHisLeuTyrTyrPhe  
TGCATACAGCAAGCAGACTGCCATCTCGGCTCGGCACTGTCATAGAAATGAGGAAAGAGATATAGCACACAGTAGACCTGAAGTACGACAGCAATTAATTCATCTGTATTACT  
4800

AspCysThrSerAspSerAlaIleAlaLeuLeuGlyHisIleValTyrSerProArgCysGluTrpGlnAlaGlyHisAlaLysValGlnLysLeuLeuLeuAlaAla  
TACTGCTTTTTCAGCTCTGCTATGAAGAGGCTCTATTAGCAGATATAGTATAGCCTAGCTAGCTGCTGAATATACAGCAGCAGATACAAAGCTAGGCTCTCTCAATCTGCTGCTACGAC  
4900

LeulleThrProLysLysIleLysProProLeuProSerValThrLysLeuThrGluAspArgTrpAsnLysProGlnLysThrLysGlyHisArgGlySerHisThrMetAsnGlyHis  
CATTAAATACACAAAAAGATAAAGCCACCTTTCGCTAGTGTATGCAAGTACGACAGCATAGTGTGAACAGCCCGCAGAGCAGCAGGCCACAGGGAGCCACCAATTAATGAC  
5000

ACTAGAGCTTTTACAGGAGCTTAAGATCAAGCTGTGAGACATTTTCTAGGATTTGCTCCATGGCTTAGGCCACATATCTATGAAGTTATGGGATACTTGGCCAGGAGTGCAGCC  
5100

CATAATAGCAATTTGCAACCACTGCTGTTTATGCTATTCAGAAATGGGTGTCGACATAGCAAAATAGGCGTTACTCAACAGAGGAGAGCAAAATGGAGCCAGTACATCTAGACTAG  
5200

AGCGCTGCAAGCATCGAGGAGTACGCTAAAAGTGGTGTAGCAGTTCGTAATGTAAGAAAGTGTGCTTTTCAATGCGCAAGTTTGTTCACAAAAAGGCTTACGCACTCTCTGTCGCA  
5300

GCAAGCAGCGGACACCGCAGGAAGCTCTCTCAAGCACTCAGACTCATCAAGTTTCTATCAAGCAGTAACTACTACATGTAATGCAACCTATACAAATAGCAATAGCAGCATAG  
5400

5500

5600

5700

5800

5900

6000

6100

6200

6300

6400

6500

6600

6700

6800

6900

7000

7100

7200

7300

7400

7500

7600

7700

7800

7900

8000

8100

8200

8300

8400

8500

8600

8700

8800

8900

9000

9100

9200

9300

9400

9500

9600

9700

9800

9900

[illegible]

**Figure 1. Complete DNA Sequence of Viral Genome (LAV-1a)**

The sequence was reconstructed from the sequence of phage  $\lambda$ 119 insert. The numbering starts at the cap site, which was located experimentally (see above). Important genetic elements, major open reading frames, and their predicted products are indicated together with the Hind III cloning sites. The potential glycosylation sites in the *env* gene are overlined. The NH<sub>2</sub>-terminal sequence of p25<sup>99</sup> determined by protein microsequencing is boxed (Genetic Systems, personal communication).

Each nucleotide was sequenced on average 5.3 times: 85% of the sequence was determined on both strands and the remainder was sequenced at least twice from independent clones. The base composition is T, 22.2%; C, 17.8%; A, 35.8%; G, 24.2%; G + C, 42%. The dinucleotide CpG is greatly under-represented (0.9%) as is common among eukaryotic sequences (Bird, 1980).

## The LTR

The organization of a reconstructed LTR and viral flanking elements are shown schematically in Figure 3. The LTR is 638 bp long and displays usual features (Chen and Barker, 1984): it is bounded by an inverted repeat (5'ACTG) including the conserved TG dinucleotide (Temin, 1981); adjacent to 5' LTR is the tRNA primer binding site (PBS), complementary to tRNA<sup>Lys</sup> (Raba et al., 1979); adjacent to 3' LTR is a perfect 15 bp polypurine tract. The other three

polypurine tracts observed between nucleotides 8200-8800 are not followed by a sequence that is complementary to that just preceding the PBS.

The limits of U5, R, and U3 elements were determined as follows. U5 is located between PBS and the polyadenylation site established from the sequence of the 3' end of oligo(dT)-primed LAV cDNA (Alizón et al., 1984). Thus U5 is 84 bp long. The length of R+U5 was determined by synthesizing tRNA-primed LAV cDNA. After alkaline hydroly-

Table 1. Locations and Sizes of Viral Open Reading Frames

orf	1 <sup>st</sup> Triplet	Met	Stop	No. Amino Acids	M <sub>r</sub> Calc.
gag	312	336	1,836	500	55,841
pol	1,631	1,934	4,640	(1,003)	(113,629)
orf Q	4,554	4,587	5,163	192	22,487
env	5,746	5,767	8,350	861	97,376
orf F	8,324	8,354	8,972	206	23,316

The nucleotide coordinates refer to the first base of the first triplet (1<sup>st</sup> triplet), of the first methionine (initiation) codon (Met) and of the stop codon (Stop). The numbers of amino acids and molecular weights are those calculated for unmodified precursor products starting at the first methionine through to the end, with the exception of pol, where the size and M<sub>r</sub> refer to that of the whole orf.

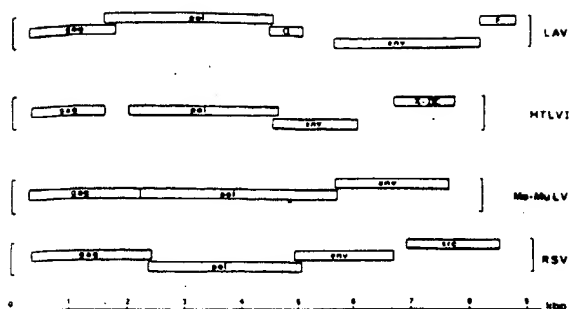


Figure 2. Comparison of the Genome Organization of LAV with Those of Human T Cell Leukemia/Lymphoma Virus Type I (HTLV-I) (Seiki et al., 1983), Moloney Murine Leukemia Virus (MoMuLV) (Shinnick et al., 1981), and Rous Sarcoma Virus (RSV) (Schwartz et al., 1983)

The positions and sizes of viral genes are drawn to scale (open boxes) and the viral genomes (RNA forms) are delimited by brackets.

sis of the primer, R+U5 was found to be  $181 \pm 1$  bp (Figure 4). Thus R is 97 bp long and the cap site at its 5' end can be located. Finally, U3 is 456 bp long. The LAV LTR also contains characteristic regulatory elements: a polyadenylation signal sequence AATAAA 19 bp from the R-U5 junction, and the sequence ATATAAG, which is very likely the TATA box, 22 bp 5' of the cap site. There are no long direct repeats within the LTR. Interestingly, the LAV LTR shows some similarities to that of the mouse mammary tumor virus (MMTV) (Donehower et al., 1981). They both use tRNA<sup>phe</sup> as a primer for (-) strand synthesis, whereas all other exogenous mammalian retroviruses known to date use tRNA<sup>pro</sup> (Chen and Barker, 1984). They possess very similar polypurine tracts; that of LAV is AAAAGAAAAGGGGGG while that of MMTV is AAAAAAGAAAAAGGGGGG. It is probable that the viral (+) strand synthesis is discontinuous since the polypurine tract flanking the U3 element of the 3'LTR is found exactly duplicated in the 3' end of orf pol, at 4331-4346. In addition, MMTV and LAV are exceptional in that the U3 element can encode an orf. In the case of MMTV, U3 contains the whole orf while, in LAV, U3 contains 110 codons of the 3' half of orf F.

### Viral Proteins

#### gag

Near the 5' extremity of the gag orf is a "typical" initiation codon (Kozak, 1984) (position 336), which is not only the first in the gag orf, but the first from the cap site. The precursor protein is 500 amino acids long. The calculated  $M_r$  of 55,841 agrees with the 55 kd gag precursor polypeptide (Luc Montagnier, unpublished results). The N-terminal amino acid sequence of the major core protein p25, obtained by microsequencing (Genetic Systems, personal communication), matches perfectly with the translated nucleotide sequence starting from position 732 (see Figure 1). This formally makes the link between the cloned LAV genome and the immunologically characterized LAV p25 protein. The protein encoded 5' of the p25 coding sequence is rather hydrophilic. Its calculated  $M_r$  of 14,866 is consistent with that of the gag protein p18. The 3' part of the gag region probably codes for the retroviral nucleic acid binding protein (NBP). Indeed, as in HTLV-I (Seiki et

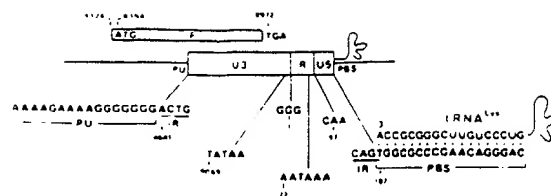


Figure 3. Schematic Representation of the LAV Long Terminal Repeat (LTR)

The LTR was reconstructed from the sequence of LAV by juxtaposing the sequences adjacent to the Hind III cloning sites. Sequencing of oligo(dT)-primed LAV DNA clone pLAV75 (Alizon et al., 1984) rules out the possibility of clustered Hind III sites in the R region of LAV. LTR are limited by an inverted repeat sequence (IR). Both of the viral elements flanking the LTR have been represented as tRNA primer binding site (PBS) for 5' LTR and polypurine track (PU) for 3' LTR. Also indicated are a putative TATA box, the cap site, polyadenylation signal (AATAAA), and polyadenylation site (CAA). The location of the open reading frame F (648 nucleotides) is shown above the LTR scheme.

al., 1983) and RSV (Schwartz et al., 1983), the motif Cys-X<sub>2</sub>-Cys-X<sub>2</sub>-Cys common to all NBP (Oroszian et al., 1984) rules out the possibility of clustered Hind III sites in the R region of LAV. LTR are limited by an inverted repeat sequence (IR). Both of the viral elements flanking the LTR have been represented as tRNA primer binding site (PBS) for 5' LTR and polypurine track (PU) for 3' LTR. Also indicated are a putative TATA box, the cap site, polyadenylation signal (AATAAA), and polyadenylation site (CAA). The location of the open reading frame F (648 nucleotides) is shown above the LTR scheme.

#### pol

The reverse transcriptase gene can encode a protein of up to 1003 amino acids (calculated  $M_r = 113,629$ ). Since the first methionine codon is 92 triplets from the origin of the open reading frame, it is possible that the protein is translated from a spliced messenger RNA, giving a gag-pol polypeptide precursor.

The pol coding region is the only one in which significant homology has been found with other retroviral protein sequences, three domains of homology being apparent. The first is a very short region of 17 amino acids (starting at 1856). Homologous regions are located within the p15 gag<sup>RSV</sup> protease (Dittmar and Moelling, 1978) and a polypeptide encoded by an open reading frame located between gag and pol of HTLV-I (Figure 5) (Schwartz et al., 1983; Seiki et al., 1983). This first domain could thus correspond to a conserved sequence in viral proteases. Its different locations within the three genomes may not be significant since retroviruses, by splicing or other mechanisms, express a gag-pol polypeptide precursor (Schwartz et al., 1983; Seiki et al., 1983). The second and most extensive region of homology (starting at 2048) probably represents the core sequence of the reverse transcriptase. Over a region of 250 amino acids, with only minimal insertions or deletions, LAV shows 38% amino acid identity with RSV, 25% with HTLV-I, and 21% with MoMuLV (Schinnick et al., 1981) while HTLV-I and RSV show 38% identity in the same region. A third homologous region is situated at the 3' end of the pol reading frame and corresponds to part of the pp32 peptide of RSV that has exonuclease activity (Misra et al., 1982). Once again, there is greater homology with the corresponding RSV sequence than with HTLV-I.

#### env

The env open reading frame has a possible initiator methionine codon very near the beginning (eighteenth codon).

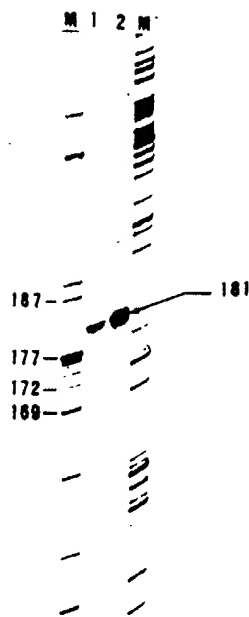


Figure 4. Synthesis of RNA-Primed LAV cDNA for R+U5 (Strong-Stop cDNA)

Lanes 1 and 2 show two different quantities of cDNA while lanes M and M' represent markers. The strong-stop cDNA is 181 bases long with a second, less intense band at 180. The error of estimation is  $\pm 1$  bp. This maps the major cap site to the second G residue of the sequence CTGGGTCT within the LTR, 24 nucleotides downstream of the TATA box. This guanosine residue is taken as the first base in the nucleotide sequence shown in Figure 1.

If so, the molecular weight of the presumed env precursor protein (861 amino acids,  $M_r$  calc = 97,376) is consistent with the known size of the LAV glycoprotein (110 kd and 90 kd after glycosidase treatment; Luc Montagnier, unpublished). There are 32 potential N-glycosylation sites (Asn-X-Ser/Thr), which are overlined in Figure 1. An interesting feature of env is the very high number of Trp residues at both ends of the protein. There are three hydrophobic regions, characteristic of the retroviral envelope proteins (Seiki et al., 1983), corresponding to a signal peptide (encoded by nucleotides 5815–5850 bp), a second region (7315–7350 bp), and a transmembrane segment (7831–7896 bp). The second hydrophobic region (7315–7350 bp) is preceded by a stretch rich in Arg + Lys. It is possible that this represents a site of proteolytic cleavage, which, by analogy with other retroviral proteins, would give an external envelope polypeptide and a membrane-associated protein (Seiki et al., 1983; Kiyokawa et al., 1984). A striking feature of the LAV envelope protein sequence is that the region following the transmembrane segment is of unusual length (150 residues). The env protein shows no

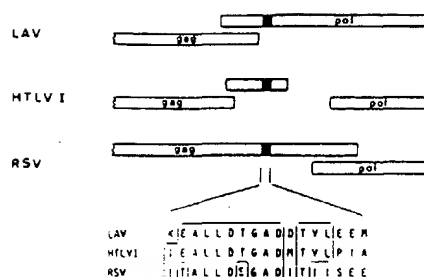


Figure 5. Location of a Short Stretch of Homology in the gag-pol Region of the LAV, HTLV-I (Seiki et al., 1983) and RSV (Schwartz et al., 1983) Genomes

Conserved amino acids are boxed. Homologous region is shown by the solid bar in the schema. Each virus is organized differently in this region but the sequence in the RSV genome maps to p15<sup>gag</sup>, which has a protease-associated function.

homology to any sequence in protein data banks. The small amino acid motif common to the transmembrane proteins of all leukemogenic retroviruses (Cianciolo et al., 1984) is not present in LAV env.

#### Q and F

The location of orf Q is without precedent in the structure of retroviruses. Orf F is unique in that it is half-encoded by the U3 element of the LTR. Both orf have strong initiator codons (Kozak, 1984) near their 5' ends and can encode proteins of 192 amino acids ( $M_r$  calc = 22,487) and 206 amino acids ( $M_r$  calc = 23,316), respectively. Both putative proteins are hydrophilic (pQ 49% polar, 15.1% Arg + Lys; pF 46% polar, 11% Arg + Lys) and are therefore unlikely to be associated directly with membrane. The function for the putative proteins pQ and pF cannot be predicted, as no homology was found by screening protein sequence data banks. Between orf F and the pX protein of HTLV-I there is no detectable homology. Furthermore, their hydrophobicity/hydrophilicity profiles are completely different. It is known that retroviruses can transduce cellular genes—notably proto-oncogenes (Weinberg, 1982). We suggest that orfs Q and F represent exogenous genetic material and not some vestige of cellular DNA because LAV DNA does not hybridize to the human genome under stringent conditions (Alizon et al., 1984), and their codon usage is comparable to that of the gag, pol, and env genes (data not shown).

#### Relationship to Other Retroviruses

Although LAV is both morphologically and biochemically (Barré-Sinoussi et al., 1983) distinct to HTLV-I and -II, it remained possible that its genome was organized in a similar manner. The characteristic features of HTLV-I and -II genomes, which they share with the more distantly related bovine leukemia virus (BLV) (Rice et al., 1984), are not observed in the case of LAV. These are: a region 3' of the envelope gene consisting of a noncoding stretch (600–900 bp), followed by a coding sequence of 307–357 codons (X open reading frame), which may slightly overlap the U3 region of the LTR (Seiki et al., 1983; Rice et al., 1984; Sagata et al., 1984) and, second, the LTR being



Table 2. Comparison of the Size of the LAV LTR and LTR-Related Element to Those of Other Retroviruses

	LTR	U3	R	U5	PU	PBS	IR
LAV	638	456	97	85	15	LYS	4
HTLV-I	759	355	228	176	12'	PRO	4'
HTLV-II	763	314	248	261	12'	PRO	4'
MMTV	1,332	1,197	11	124	19	LYS	8'
MoMuLV	594	449	68	77	13	PRO	13
RSV	335	234	21	80	11	TRP	15
SNV	601	420	97	80	13	PRO	9

Adapted from Chen and Barker (1984).

i = imperfect match or tract.

SNV = spleen necrosis virus (Shimotohno and Temin, 1982).

composed of unusually long U5 and R elements and the polyadenylation signal being situated in U3 instead of R (Seiki et al., 1983; Sagata et al., 1984; Shimotohno et al., 1984). We show here that, in contrast, the 3' end of the LAV envelope gene overlaps an open reading frame, termed F, that has the coding capacity for 206 amino acids and extends within the LTR (110 amino acids are encoded by the U3 region). The putatively encoded polypeptide (pF), the primary structure of which can be deduced, does not show any homology with the theoretical X gene products of the HTLV/BLV family. Also, the U5 and R elements are shorter (Table 2) and the polyadenylation signal is located within R, as is the case for all retroviruses except the HTLV/BLV. Additionally, LAV uses tRNA<sup>lys</sup> as (-) strand primer, as opposed to tRNA<sup>pro</sup> employed by all other mammalian retroviruses except MMTV (Donehower et al., 1981). Those homologies detected between the polymerase and protease domains of LAV and HTLV are also found in several retroviruses, RSV in particular.

It has been reported that a cloned HTLV-III genome hybridizes ( $T_m = 28^\circ\text{C}$ ) to sequences in the gag-pol and X regions of HTLV-I and -II; although restriction maps of cloned LAV and HTLV-III show almost perfect agreement (Hahn et al., 1984), we were unable to detect any such hybridization between LAV and HTLV-II ( $T_m = 55^\circ\text{C}$ ) (Alizon et al., 1984). Indeed, there is a punctual region of homology between LAV and HTLV-I (23/27 nucleotides starting at position 1859 in the LAV sequence) but nothing significant between the two viruses in the X region of HTLV-I. One possible reason for this discrepancy is that HTLV-III is subtly different from LAV. However it was subsequently reported that there was very minimal, if any, homology between *orf* X (of HTLV-I) and HTLV-III (Shaw et al., 1984).

## Discussion

Regulatory sequences carried by retroviral LTR are believed to be involved in specific interactions between the viral genome and the host cell (Srinivasan et al., 1984). The LTR sequences of LAV are unique among retroviruses. That could reflect an original mode of gene expression, possibly in relation to particular transcriptional factors present in the virus-harboring cell. This hypothesis can be tested by studying the regulatory activity of the LAV

LTR sequences in transient or long-term experiments involving an indicator gene and different cellular contexts.

The presence of the Q and F reading frames in addition to the conventional gag-pol-env set of genes is unexpected. One should now address the question of their role in the viral cycle and pathogenicity by trying to characterize their protein product(s). It is tempting to speculate on a role of such polypeptide(s) in T4 cells' mortality, a problem that can be studied by designing synthetic peptides for antibody production or by using site-directed mutagenesis of Q and F coding regions.

The peculiar genetic structure of LAV poses the question of its origin. The virus shares common tracts with other (apparently unrelated) retroviruses. For instance, the unusually large size of the outer membrane glycoprotein (env) and a comparably sized genome are also observed in the case of lentiviruses such as Visna (Harris et al., 1981; Quérat et al., 1984). The presence of a large part of the F open reading frame in the LTR, and the use of tRNA<sup>lys</sup> as a primer for (-) strand synthesis, is reminiscent of the mouse mammary tumor virus. On the other hand, homologies in the pol gene would suggest that the LAV is closer to RSV than to any other retroviruses. Obviously, no clear picture can be drawn from the DNA sequence analysis as far as phylogeny is concerned. Thus, it may well be that LAV defines a new group of retroviruses that have been independently evolving for a considerable period of time, and not simply a variant recently derived from a characterized viral family. Both epidemiology and pathogeny of AIDS should be reconsidered with this idea in mind, when trying to answer such questions as these: Are there other human or animal diseases that are associated with similarly organized viruses? Is there a precursor to AIDS-associated virus(es) normally present, in latent form, in human populations? What triggered in this case the recent spreading of pathogenic derivatives?

## Experimental Procedures

### M13 Cloning and Sequencing

Total  $\lambda$ 19 DNA was sonicated, treated with the Klenow fragment of DNA polymerase plus deoxynucleotides (2 hr,  $16^\circ\text{C}$ ), and fractionated by agarose gel electrophoresis. Fragments of 300–600 bp were excised, electroeluted, and purified by Elutip (Schleicher and Schüll) chromatography. DNA was ethanol-precipitated using 10  $\mu\text{g}$  dextran T40 (Pharmacia) as carrier and ligated to dephosphorylated, Sma I-cleaved M13mp8 RF DNA using T4 DNA and RNA ligases (16 hr,  $16^\circ\text{C}$ ) and transfected into *E. coli* strain TG-1. Recombinant clones were detected by plaque hybridization using the appropriate <sup>32</sup>P-labeled LAV restriction fragments as probes. Single-stranded templates were prepared from plaques exhibiting positive hybridization signals and were sequenced by the dideoxy chain termination procedure (Sanger et al., 1977) using  $\alpha$ -<sup>32</sup>S-dATP (Amersham, 400 Ci/mmol) and buffer gradient gels (Biggen et al., 1983). Sequences were compiled and analyzed using the programs of Staden adapted by B. Caudron for the Institut Pasteur Computer Center (Staden, 1982).

### Strong-Stop cDNA

LAV virions from infected T lymphocyte (Barré-Sinoussi et al., 1983) culture supernatant were pelleted through a 20% sucrose cushion and the cDNA (-) strand was synthesized as described previously (Alizon et al., 1984) except that no exogenous primer was used. After alkaline hydrolysis (0.3 M NaOH, 30 min,  $65^\circ\text{C}$ ), neutralization, and phenol extraction, the cDNA was ethanol-precipitated and loaded onto a 5%

acrylamide/8 M urea sequencing gel with sequence ladders as size markers.

#### Acknowledgments

We would like to thank Professors Luc Montagnier and Pierre Tiollais, in whose laboratory this work was carried out, for support and encouragement, as well as Professor Raymond Dedonder and Agnes Ullmann for their commitment to the project. Bernard Caudron and Jean-Noël Paulous of the Institut Pasteur Computer Center provided invaluable and constant assistance, and Michelle Fonck, technical support. Ana Cova and Louise-Marie Da tirelessly and good-humoredly typed the manuscript. We would like to thank Dr. Moshe Yaniv for critical reading of the manuscript and, finally, Genetic Systems, Seattle, WA, for communicating unpublished data.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received December 26, 1984

#### References

- Alizon, M., Sonigo, P., Barré-Sinoussi, F., Chermann, J. C., Tiollais, P., Montagnier, L., and Wain-Hobson, S. (1984). Molecular cloning of lymphadenopathy-associated virus. *Nature*, in press.
- Arya, S. K., Gallo, R. C., Hahn, B. H., Shaw, G. M., Popovic, M., Salathuddin, S. Z., and Wong-Staal, F. (1984). Homology of genome of AIDS-associated virus with genomes of human T-cell leukemia lymphoma viruses. *Science* 225, 927-930.
- Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Daugey, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk of acquired immune deficiency syndrome (AIDS). *Science* 220, 868-870.
- Biggen, M. D., Gibson, T. J., and Hong, G. F. (1983). Buffer gradient gels and <sup>32</sup>S label as an aid to rapid DNA sequence determination. *Proc. Natl. Acad. Sci. USA* 80, 3963-3965.
- Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucl. Acids Res.* 8, 1499-1504.
- Brun-Vézinet, F., Rouzioux, C., Barré-Sinoussi, F., Klatzmann, D., Saimot, A. G., Rozenbaum, W., Montagnier, L., and Chermann, J. C. (1984). Detection of IgG antibodies to lymphadenopathy associated virus (LAV) by ELISA, in patients with acquired immunodeficiency syndrome of lymphadenopathy syndrome. *Lancet* i, 1253-1256.
- Chen, H. R., and Barker, W. C. (1984). Nucleotide sequences of the retroviral long terminal repeats and their adjacent regions. *Nucl. Acids Res.* 12, 1767-1778.
- Chen, I. S. Y., McLaughlin, J., Gasson, J. C., Clark, S. C., and Golde, D. W. (1983). Molecular characterization of the genome of a novel human T-cell leukaemia virus. *Nature* 305, 502-505.
- Chiu, I. M., Callahan, R., Tronick, S. R., Scholm, J., and Aaronson, S. A. (1984). Major pol gene progenitors in the evolution of oncoviruses. *Science* 223, 364-370.
- Cianciolo, G. J., Kipnis, R. J., and Snyderman, R. (1984). Similarity between p15E of murine and feline viruses and p21 of HTLV. *Nature* 311, 515.
- Daly, H. M., and Scott, G. L. (1983). Fatal AIDS in a haemophiliac in the U.K. *Lancet* ii, 1190.
- Dittmar, K. J., and Moelling, K. (1978). Biochemical properties of p15-associated protease in an avian RNA tumor virus. *J. Virol.* 28, 106-118.
- Donehower, L. A., Huang, A. L., and Hager, G. L. (1981). Regulatory and coding potential of the mouse mammary tumour virus long terminal redundancy. *J. Virol.* 37, 226-238.
- Gottlieb, M. S., Schroff, R., Schanler, H. M., Weisman, J. D., Fan, P. T., Wolf, R. A., and Saxon, A. (1981). Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *N. Eng. J. Med.* 305, 1426-1431.
- Hahn, B. H., Shaw, G. M., Arya, S. U., Popovic, M., Gallo, R. C., and Wong-Staal, F. (1984). Molecular cloning and characterization of the HTLV-III virus associated with AIDS. *Nature* 312, 166-169.
- Harris, J. D., Scott, J. V., Taylor, B., Brahic, M., Stowring, L., Ventura, P., Haase, A. T., and Peluso, R. (1981). Visna virus DNA: discovery of a novel gapped structure. *Virology* 113, 573-583.
- Kiyokawa, T., Yoshikura, H., Hattori, S., Secki, M., and Yoshida, M. (1984). Envelope proteins of human T-cell leukemia virus: expression in Escherichia coli and its application to studies of env gene functions. *Proc. Natl. Acad. Sci. USA* 81, 6202-6206.
- Klatzmann, D., Barré-Sinoussi, F., Nugeyre, M. T., Daugey, C., Vilmer, E., Griscelli, C., Brun-Vézinet, F., Rouzioux, C., Gluckman, J. C., Chermann, J. C., and Montagnier, L. (1984). Selective tropism of lymphadenopathy associated virus (LAV) for helper-inducer T-lymphocytes. *Science* 225, 59-63.
- Kozak, M. (1984). Compilation and analysis of sequences upstream from the transcriptional start site in eucaryotic mRNAs. *Nucl. Acids Res.* 12, 857-872.
- Levy, J. A., Hoffman, A. D., Kramer, S. M., Lanois, J. A., Shimabukuro, J. M., and Oskiro, L. S. (1984). Isolation of lymphocytotropic retroviruses from San Francisco patients with AIDS. *Science* 225, 840-842.
- Masur, H., Michelis, M. A., Greene, J. B., Onovato, I., Van de Stowe, R. A., Holzman, R. S., Wormser, G., Brettman, L., Lange, M., Murray, H. W., and Cunningham-Rundles, S. (1981). An outbreak of community-acquired pneumocystis carinii pneumonia: initial manifestation of cellular immune dysfunction. *N. Eng. J. Med.* 305, 1431-1438.
- Misra, T. K., Grandgenett, D. P., and Parsons, J. T. (1982). Avian retrovirus pp32 DNA-binding protein. I. Recognition of specific sequences on retrovirus DNA terminal repeats. *J. Virol.* 44, 330-343.
- Montagnier, L., Chermann, J. C., Barré-Sinoussi, F., Chamaret, S., Gruest, J., Nugeyre, M. T., Rey, F., Daugey, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Saimot, A. G., Rozenbaum, W., Gluckman, J. C., Klatzmann, D., Vilmer, E., Griscelli, C., Gazengel, C., and Brunet, J. B. (1984). A new human T-lymphotropic retrovirus: characterization and possible role in lymphadenopathy and acquired immune deficiency syndromes. In *Human T-Cell Leukemia/Lymphoma Virus*, R. C. Gallo, M. Essex, and L. Gross, eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory), pp. 363-370.
- Oroszlan, S., Copeland, T. D., Kalyanaraman, V. S., Sarngadharan, M. G., Schultz, A. M., and Gallo, R. C. (1984). Chemical analysis of human T-cell leukemia virus structural proteins. In *Human T-Cell Leukemia/Lymphoma Virus*, R. C. Gallo, M. E. Essex, and L. Gross, eds. (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory), pp. 101-110.
- Piot, P., Quinn, T. C., Taelmann, H., Feinsod, F. M., Minlangu, K. B., Wobin, O., Mbendi, N., Mazabo, P., Ndangi, K., Stevens, W., Kalam-bayi, K., Mitchell, S., Bridts, C., and McCormick, J. B. (1984). Acquired immunodeficiency syndrome in a heterosexual population in Zaire. *Lancet* ii, 65-69.
- Popovic, M., Sarngadharan, M. G., Read, E., and Gallo, R. C. (1984). Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science* 224, 497-500.
- Querat, G., Barban, N., Sauze, N., Filippi, P., Vigne, R., Russo, P., and Vitu, C. (1984). Highly lytic and persistent lentiviruses naturally present in sheep with progressive pneumonia are genetically distinct. *J. Virol.* 52, 672-679.
- Raba, M., Limburg, K., Burghagen, M., Katze, J. R., Simsek, M., Heckman, J. E., Rajbhandary, U. L., and Gross, H. J. (1979). Nucleotide sequence of three isoaccepting lysine tRNAs from rabbit liver and SV40-transformed mouse fibroblasts. *Eur. J. Biochem.* 97, 305-318.
- Rice, N. R., Stephen, R. M., Couez, D., Deschamps, J., Kettmann, R., Burny, A., and Gilden, R. V. (1984). The nucleotide sequence of the env gene and post-env region of bovine leukemia virus. *Virology* 138, 82-93.
- Sagata, N., Yasunaga, T., Ogawa, Y., Tsuzuku-Kawamura, J., and Ikawa, Y. (1984). Bovine leukemia virus: unique structural features of its long terminal repeats and its evolutionary relationship to human T-cell leukemia virus. *Proc. Natl. Acad. Sci. USA* 81, 4741-4745.
- Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing

with chain terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.

Schüpbach, J., Popovic, M., Gilden, R. V., Gonda, M. A., Sarngadharan, M. G., and Gallo, R. C. (1984). Serological analysis of a subgroup of human T-lymphotropic retroviruses (HTLV-III) associated with AIDS. *Science* 224, 503-505.

Schwartz, D. E., Tizard, R., and Gilbert, W. (1983). Nucleotide sequence of Rous sarcoma virus. *Cell* 32, 853-869.

Seiki, M., Hattori, S., Hirayama, Y., and Yoshida, M. (1983). Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. *Proc. Natl. Acad. Sci. USA* 80, 3618-3622.

Shaw, G. M., Hahn, B. H., Arya, S. K., Groopman, J. E., Gallo, R. C., and Wong-Staal, F. (1984). Molecular characterization of human T-cell leukemia (lymphotropic) virus type III in the acquired immune deficiency syndrome. *Science* 226, 1165-1171.

Shimotohno, K., and Temin, H. M. (1982). Spontaneous variation and synthesis in the U3 region of the long terminal repeat of an avian retrovirus. *J. Virol.* 41, 163-171.

Shimotohno, K., Golde, D. M., Miwa, M., Sugimura, T., and Chen, I. S. Y. (1984). Nucleotide sequence analysis of the long terminal repeat of human T-cell leukemia virus type II. *Proc. Natl. Acad. Sci. USA* 81, 1079-1083.

Shinnick, T. M., Lerner, R. A., and Sutcliffe, J. G. (1981). Nucleotide sequence of Moloney murine leukemia virus. *Nature* 293, 543-548.

Srinivasan, A., Reddy, E. P., Dunn, C. Y., and Aaronson, S. A. (1984). Molecular dissection of transcriptional control elements with the long terminal repeat of retrovirus. *Science* 223, 286-289.

Staden, R. (1982). Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucl. Acids. Res.* 10, 4731-4751.

Temin, H. (1981). Structure, variation and synthesis of retrovirus long terminal repeat. *Cell* 27, 1-3.

Weinberg, R. A. (1982). Fewer and fewer oncogenes. *Cell* 30, 3-9.

## Complete Nucleotide Sequences of Functional Clones of the AIDS Virus

LEE RATNER,<sup>1</sup> AMANDA FISHER,<sup>2</sup> LINDA L. JAGODZINSKI,<sup>3</sup> HIROAKI MITSUYA,<sup>4</sup>  
RUEY-SHYAN LIOU,<sup>3</sup> ROBERT C. GALLO,<sup>2</sup> and FLOSSIE WONG-STAAAL<sup>2</sup>

<sup>1</sup>*Division of Hematology and Oncology, Departments of Medicine and Microbiology and Immunology,  
Washington University, St. Louis, MO*

<sup>2</sup>*Laboratory of Tumor Cell Biology, National Cancer Institute, National Institutes of Health, Bethesda, MD*

<sup>3</sup>*Biotech Research Laboratories, Inc., Rockville, MD*

<sup>4</sup>*Clinical Oncology Program, National Cancer Institute, National Institutes of Health, Bethesda, MD*

### ABSTRACT

To examine the mechanism of lymphocytotoxicity induced by human T-lymphotropic virus type III/lymphadenopathy associated virus (HTLV-III/LAV), an in vitro model has been developed. Introduction of an HTLV-III/LAV proviral clone, HXB2, into normal lymphocytes results in the production of virions and cell death. The complete nucleotide sequence of the proviral form of HXB2 has now been determined. Its structure is quite similar to that previously determined for HTLV-III/LAV clones whose biological capacities had not previously been demonstrated. The biological function of two additional clones of HTLV-III/LAV, BH10 and HXB3, are reported. Clone BH10 which lacks the 5' long terminal repeat sequences (LTR) and a portion of the 3' LTR is reconstituted by substituting the corresponding sequences of HXB2 and is shown to be capable of generating infectious cytopathic virions. Clone HXB3, which has been partially sequenced, is also found to be capable of producing lymphocytopathic virus. Clone HXB3 differs from HXB2 in its lack of a termination codon in 3'orf, demonstrating that 3'orf plays no major role in virus replication or cytopathic activity. These data provide the necessary background to allow the identification of viral determinants of replication, cytopathic activity, and antigenicity using these functional proviral clones.

### INTRODUCTION

AIDS is a devastating illness occurring as an epidemic with more than 20,000 cases identified thus far (1,2). It represents the most severe clinical manifestation of infection by HTLV-III/LAV, also designated AIDS-related virus, ARV (3) or human immunodeficiency virus, HIV (4), which is present in 1.0-2.0 million individuals in the United States alone (5,6). There is overwhelming evidence that HTLV-III/LAV is the etiological agent in AIDS and AIDS-related syndromes (7-17). One of the most convincing lines of evidence is the recapitulation of T4 cell depletion in vitro as a result of HTLV-III/LAV infection (18-20). Thus, an understanding of the mechanism(s) involved in lymphocytopathic effects is paramount for understanding the pathogenesis of this disease.

A number of different recombinant DNA clones of HTLV-III/LAV have been obtained and analyzed by restriction enzyme digestion and/or nucleotide sequencing (21-32). This work has demonstrated that the viral genome is 9182-9213 nucleotides in length, with LTRs of 636-637 nucleotides, and at least seven genes. Three replicative genes include *gag*, *pol*, and *env* which are similar to those in other retroviruses, though *env* is longer than that of other retroviruses (28,33). A fourth gene, designated *tat*, is structurally distinct from that of other retroviruses, and encodes a trans-acting factor capable of enhancing virus expression in a positive feedback manner (34-40). A fifth gene has recently been identified, and has been named *art* or anti-repressor of transactivation (41) or *trs* or trans-repressor of splicing (42). Two additional genes, designated short open reading frame (*sor*) and 3' open reading frame (*3'orf*) are also unique to HTLV-III/LAV, but the functions of their gene products are unknown (27-29,31,43-45). An additional open reading frame, designated *R*, is also presumed to encode a protein product based on the finding of antibodies in infected individuals reactive to these sequences expressed in *E. coli* (our unpublished observations with J. Ghayeb).

To define the functions of viral proteins and locate the sequences encoding the cytopathic factor(s), an in vitro model has been established (46). Cloned viral DNA sequences are introduced into normal lymphocytes derived from umbilical cord blood using the protoplast fusion technique. Viral DNA, RNA, and proteins are readily detectable after 7-24 days, as well as virions morphologically identical to those arising from natural infection. Most notably, cell death occurs 18-30 days after transfection. Thus, transfection of HTLV-III/LAV proviral DNA into normal lymphocytes results in the production of lymphocytotropic virus, reproducing the major manifestations of infection observed both in the laboratory and in humans (18-20,46).

This experimental system allows analysis of viral sequences required for the cytopathic activity by in vitro mutagenesis prior to transfection. To provide the necessary background for the construction of these mutants, and to gain further insight into the structure of active viral proteins, we have determined the complete nucleotide sequence of the functional HTLV-III/LAV clone HXB2. In addition, we have ligated a previously sequenced HTLV-III/LAV clone, BH10 (28), to LTRs of HXB2 and have demonstrated that this clone also gives rise to cytopathic virus.

## MATERIALS AND METHODS

### Recombinant DNA Clones

A single T4-positive cell line, H9, was inoculated with pooled blood samples of different patients with AIDS or related symptoms (20). Recombinant phage clones  $\lambda$ HXB2 and  $\lambda$ HXB3 were derived from this infected cell line by cloning integrated proviral copies with flanking cellular sequences in the *Xba* I site of phage J1 (47). A 12.5 kilobase (kb) *Hpa* I - *Xba* I fragment of  $\lambda$ HXB2 was blunt-ended with Klenow fragment of DNA polymerase I and cloned into the similarly blunt-ended *Bam* HI to *Eco* RI sites of vector SP65gpt. The resultant clone HXB2gpt2 has the HTLV-III and xanthine guanine phosphoribosyltransferase (gpt) sequences in the same transcriptional orientation. SP65gpt was constructed by ligating the *Bam* HI - *Pvu* II fragment of pSV2gpt (48) into the *Bam* HI - *Pvu* II sites of SP65 (Promega Scientific). Other subclones of  $\lambda$ HXB2 were made in SP65 and SP62 (New England Biolabs).

$\lambda$ BH10 was derived from the same infected cell line by cloning an unintegrated viral copy in the *Sst* I site of  $\lambda$ gtw10- $\lambda$ b (25). The 8933 nucleotide insert of  $\lambda$ BH10 (nucleotides 222-9154, based on the numbering scheme in ref. 28) was subcloned into the *Sst* I site of SP64 (Promega Scientific).

Plasmid HXB3 was constructed by subcloning the 13.0 kb *Xba* I insert of  $\lambda$ HXB3 in the *Xba* I site of SP62. Clone HXB2/3gpt was made by replacing the 2.3 kb *Xho* I - *Xba* I fragment of HXB2gpt with the 1.2 kb *Xho* I - *Xba* I insert of HXB3.

### DNA Seq

Th  
flankin  
method  
chain t

### Assays

Cy  
in norm  
virus t  
(46,51,  
polyeth  
mononuc  
mented  
(IL2,  
strepto  
dye exc  
3-5 day  
Pl  
into H9  
from 2-  
at 100,  
medium  
electro  
x 10<sup>4</sup> -  
brene.  
culture  
FCS, 5%  
mine, s  
At vari  
exclusi

Th  
Seventy  
quenced  
are lib  
alterat  
and 2)  
III/LAV  
shifts  
tion, t  
single  
*gag* and  
copies  
Th  
clone  
mented  
of clon  
*gag* to  
clone F  
BH10 or  
*Sst* I  
HX10gpt  
positio  
BH10 at  
cytes  
tion (F

### DNA Sequencing

The entire HXB2 proviral sequence and about 300 nucleotides of 5' and 3' flanking cellular sequences were determined by the partial chemical cleavage method (49) except nucleotides 3306-3739, which were determined by the dideoxy chain termination method (50).

### Assays for Cytopathic Activity

Cytopathic activity of viruses derived from proviral DNA clones was tested in normal umbilical cord blood mononuclear cells or a human T-lymphotropic virus type I immortalized nonproducer cell line, ATH8, as previously described (46,51,52).  $10^{10}$  protoplasts carrying a plasmid DNA sequence were fused using polyethylene glycol with  $2 \times 10^6$  phytohemagglutinin (PHA) stimulated cord blood mononuclear cells. Cultures were grown in RPMI-1640 medium (Gibco) supplemented with 10% (v/v) fetal calf serum (FCS, Gibco), 10% (v/v) interleukin 2 (IL2, lectin-depleted; Cellular Products), 50 u/ml penicillin, and 50 µg/ml streptomycin (Gibco). Viable cell counts were determined using the trypan blue dye exclusion method (53) or by examination using phase contrast microscopy, at 3-5 day intervals.

Plasmids with the HTLV-III/LAV proviral sequences were also transfected into H9 cells by the protoplast fusion method. Virus preparations were made from 2-4 liters of cell-free supernatants of these cultures by centrifugation at  $100,000 \times g$  for 1 hour at 4°C. The pellet was resuspended in RPMI-1640 medium and diluted to a concentration of  $10^{11}$  particles/ml as determined by electron microscopy. Concentrated virus was then added in 0.2 ml of RPMI to  $2 \times 10^4$  -  $2 \times 10^5$  ATH8 cells pretreated for 30 minutes at 37°C with 2 µg/ml polybrene. The virus was allowed to adsorb for 45 minutes at 37°C, and then the cultures were diluted to 2 ml with RPMI-1640 medium supplemented with 15% (v/v) FCS, 5% (v/v) IL2, 50 u/ml of penicillin, 50 µg/ml streptomycin, 4 mM L-glutamine, and 50 µM beta-mercaptoethanol. Cells were grown in Falcon 3033 tubes. At various time points, viable cell counts were determined by trypan blue dye exclusion.

### RESULTS

The complete nucleotide sequence of clone HXB2 is shown in Figure 1. Seventy-nine nucleotide substitutions are noted compared to a previously sequenced proviral clone, BH10 (28). Few if any of these sequence differences are likely to represent cloning artifacts or sequence errors since 1) these alterations were confirmed by DNA sequences from both strands of both clones, and 2) 82% of these changes are present in other previously sequenced HTLV-III/LAV clones (23,27-31). Of note is the lack of termination codons or frame-shifts within any of the previously described open reading frames. In addition, two insertions in HXB2 relative to BH10 are in noncoding regions, and a single in-frame 36 bp deletion is present in the region of overlap between the *gag* and *pol* genes. The latter alteration represents a deletion of one of two copies of a perfect tandemly repeated sequence present in BH10.

The functional capabilities of clone BH10 were also examined. Since this clone lacked the complete viral sequence, the missing portions were complemented by those obtained from clone HXB2 (Fig. 2a). The *Cla* I - *Xho* I insert of clone BH10 (nucleotides 374-8474, corresponding to amino acid residue 14 of *gag* to 44 of 3'orf) was inserted in place of that at the same positions of clone HXB2gpt2. The resultant clone HX10gpt could be distinguished from either BH10 or HXB2gpt2 by the presence of a *Bgl* II site at nucleotide position 20, a *Sst* I site at position 34, and *Hind* III sites at positions 78 and 9194 in HX10gpt and HXB2gpt2 but absent from BH10, and the presence of a *Sst* I site at position 5580 and a *Hind* III site at position 5607 in HXB2gpt2 but absent from BH10 and HX10gpt. Transfection of HX10gpt into umbilical cord blood lymphocytes resulted in virions with morphology similar to those arising with infection (Fig. 2b and 2c).

A A S S C C A C A S  
 S S S S S  
 Q Q V P Q  
  
 A A C C T T A T T G C  
 S S S S S  
 A S A C S A A S A A S  
  
 T T C T A T A A C A S  
 F S S S  
  
 A A C C C A A C C T  
 S T T A A T T G A S  
  
 T C T G T A T G T C  
 S S S S S  
 T T T T T T T G C  
 S V W A  
  
 A T A T A T A C A G  
 D I S S S  
  
 T S A T A A T G G A  
 M I M S S  
  
 A T A C T A C C A G  
 D T T T S S  
  
 A A T T T A A A  
 S C S S S  
  
 A A A G A A A G S T  
 S S S S S  
  
 T C G T A T C C A  
 I S S S  
  
 G C A A A T T A A G  
 S S S S  
  
 C A C C A C A C T  
 S S S S  
  
 G S C C A A A A G T  
 S V W  
  
 T C T T C A S A L  
 I S S S  
  
 A G A S A G A A A A  
 Q S S S  
  
 S T A T A T G T C A  
 S S S S  
  
 T A A G G A T C A  
 L E D Q  
  
 C T Y G G A T G G A  
 M S S  
  
 C A A G T T T G T A  
 A S S L W  
  
 G A T T A G G C A  
 S V W S Q  
  
 A T T A T T G A A  
 S L V S  
  
 I S S S  
  
 C C A G G G G G T  
 S R S W  
  
 S S G V  
  
 A T A G G G T T A  
 D S V  
  
 G T A T T G G A V  
 V I C  
  
 G T A C C A A T  
 A T S  
  
 T T T T A A A A  
 S S S S  
  
 G G C C A G G S  
 S S S  
  
 C T C C A T G S  
 L E G  
  
 G C T T C T A C  
 T C T C G G T  
  
 F I G U R E 1  
 T H E T R A N S  
 F O R M A T I O N  
 O F T H E A U C I  
 P R O T E I N S  
 I N T H E A U C I  
 O B L I T E R A T I O N  
 L O W E R C A S E  
 I N T H E S E S I  
 2 1 8 6 . G :  
 4 1 9 7 . A :  
 1 0 8 6 . C :  
 4 8 8 0 . T :  
 I N T H E  
 A U C I







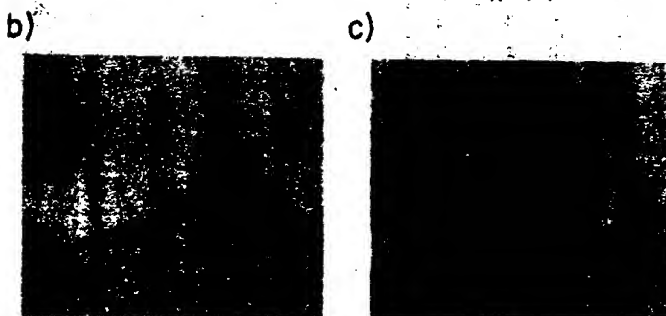
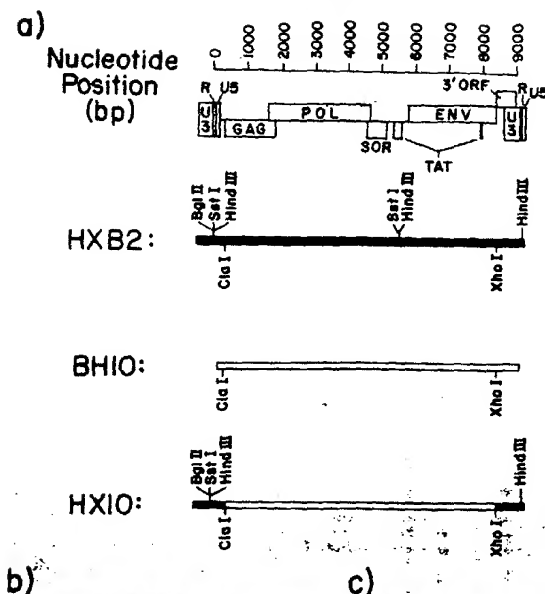


Fig. 2 - Functional capabilities of clone BH10. a) To test the activity of clone BH10 a recombinant was constructed with HXB2gpt2. The Cla I - Xho I insert of BH10 was ligated into the corresponding positions of HXB2gpt2 to generate clone HX10gpt. Bgl II, Sst I, and Hind III sites found in HX10 and/or HXB2 but not BH10 are indicated. The relative positions of these sites in the HTLV-III/LAV genome are shown by the schematic above the restriction enzyme maps. Nucleotide positions are indicated at the top. Electron micrographs of b) immature and c) mature viral particles identified 14 and 56 days, respectively, after transfection of umbilical cord mononuclear cells by protoplast fusion (46) are shown (90,000x magnification).

Clone HXB3 was also tested for its biological activity (Fig. 3). Sequences for the 3' portion of HXB3 have been determined; they differ from those of HXB2 between nucleotides 5323 and 9213 at 63 positions (23,54). It is notable that HXB3 lacks a termination codon at amino acid position 124 of 3'orf which is found in HXB2. A hybrid clone HXB2/3gpt was also constructed (Fig. 3 and Materials and Methods section), replacing the last 163 codons of 3'orf and the 3'LTR with the corresponding sequences of HXB3. Transfection of HXB3 and HXB2/3gpt into cord blood mononuclear cells or H9 cells also gave rise to infectious virions with characteristic HTLV-III/LAV morphology.

were 1  
ATH8  
SP65gpt  
cultu  
not s  
viral  
of HTI  
of HX  
killin  
that  
with  
cultu  
  
infec  
mids  
counts  
(Fig.  
testec  
moi.  
and H

1  
protei  
of the  
sequer

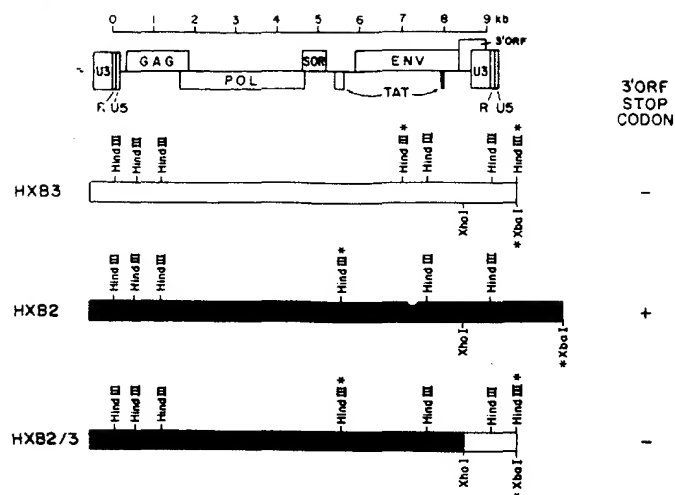


Fig. 3 - Functional HTLV-III/LAV DNA clones with or without a termination codon in 3'orf. The structure of clone HXB2 which includes a termination codon in 3'orf and clones HXB3 and HXB2/3 which lack this termination codon are shown. The positions of Hind III sites which distinguish the clones from one another are shown. In addition, the positions of Xho I and Xba I sites used for constructing the hybrid clone HXB2/3 are shown. Plasmids constructions are described in the Materials and Methods section.

Assays of the cytopathic abilities of virus produced by these DNA clones were then carried out in umbilical cord blood mononuclear cell cultures and the ATH8 cell line. Transfection into umbilical cord blood mononuclear cells of SP65gpt, which lacks HTLV-III/LAV sequences, resulted in growth of the cell culture at a rate similar to that of an untransfected culture (Fig. 4 and data not shown, ref. 46). Transfection of HXB2gpt2 resulted in the production of viral DNA, gag proteins, and viral particles with a morphology typical for that of HTLV-III/LAV, and an accelerated rate of cell death (Fig. 4). Transfection of HXB3 produced a similar rate of cell killing as did HXB2gpt2. Introduction of HX10gpt into cord blood mononuclear cells showed an attenuated rate of cell killing compared to HXB2gpt2 and HXB3, though it was reproducibly greater than that of SP65gpt. The diminished rate of cell death of cultures transfected with HX10gpt correlated with the delay in the appearance of viral compared to cultures transfected with HXB2gpt (Fig. 2, ref. 46, and data not shown).

Virus was prepared from the cell-free supernatant fluid of H9 cells either infected with an HTLV-III/LAV preparation (HTLV-IIIB) or transfected with plasmids HXB2gpt2, HX10gpt, or HXB3. These samples were then diluted by particle counts to give multiplicities of infection (moi) of 50-3000 virions per cell (Fig. 5). In each of three separate experiments, a sample with no virus was tested as well as the same reference virus preparation, HTLV-IIIB, at varying moi. These data reveal that viruses derived from DNA clones HXB2gpt2, HX10gpt, and HXB3 all produce cytopathic effects on human T cells.

## RESULTS

In order to clarify relationships between structure and function of viral proteins encoded by the HTLV-III/LAV genome, the complete nucleotide sequence of the functional clone HXB2 has been determined. In addition, the previously sequenced clone BH10 (28), when ligated to LTRs of HXB2, is also shown to be

functional. The sequence of HXB2 differs from that of BH10 in only 79 nucleotides. Furthermore, insertions of 2 and 3 nucleotides, respectively, are found in noncoding regions.

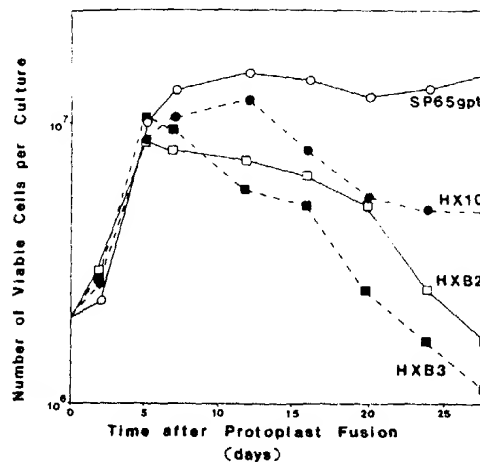


Fig. 4 - Cytopathic effects of functional HTLV-III/LAV DNA clones transfected into umbilical cord blood mononuclear cells. Plasmid clones SP65gpt, HX10gpt, HXB2gpt, and HXB3 were transfected into cord blood mononuclear cells by protoplast fusion as described in the Materials and Methods section. The number of viable cells in the cultures were determined over a period of 28 days following transfection.

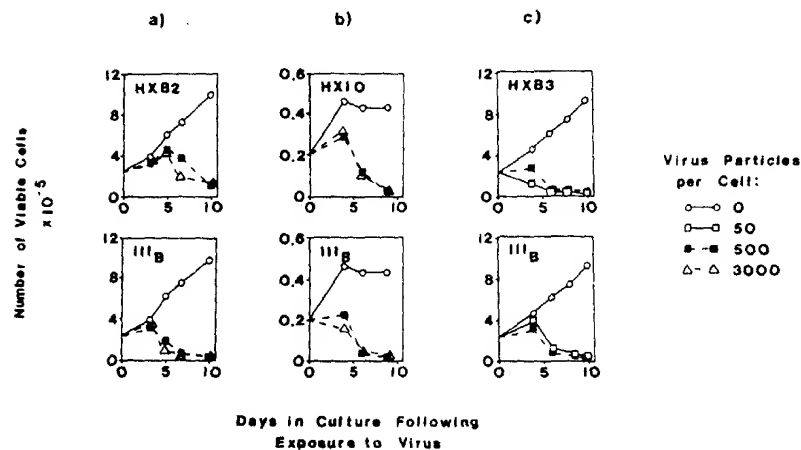


Fig. 5 - Cytopathic effects of virus derived from HTLV-III/LAV DNA clones towards ATH8 cells. Samples of virus were prepared by transfection of H9 cells with DNA clones HXB2gpt, HX10gpt, or HXB3 as described in the Materials and Methods section. Different amounts of virus were added to ATH8 cells, and the number of viable cells determined over a period of 10 days after infection. Panels a), b), and c) each represent separate experiments. The same stock virus preparation, HTLV-IIIIB (20), was used as a positive control in each case.

9 nucleo-  
are found

One notable feature of clone HXB2 is the loss of one copy of a tandemly repeated 36 bp sequence in the overlap between the gag and the pol genes, whereas most other sequenced clones, including BH10, have two copies of this sequence. This pattern of insertion or deletion of perfect or imperfect repeated DNA sequences is a common mode of variation of HTLV-III/LAV sequences, most likely occurring as a result of jumps by the reverse transcriptase during synthesis of DNA intermediates. Furthermore, the presence of one or two copies of this 36 bp sequence does not have a major influence over the rate of virus replication, and, thus, it does not perturb frameshifting which most likely occurs near this region in the synthesis of the gag-pol precursor protein.

The similarity of the structure of the functional clone HXB2 to that of BH10, and the demonstration that clone BH10 is functional, reaffirms interpretations based on the BH10 sequence data (28) and subsequent analysis of cDNA clones (27,35). The multiple open reading frame identified in the viral genome of previously sequenced clones are also found in the functional clone HXB2 are similar in size and position. Furthermore, no additional open reading frames are found in HXB2 which are absent from the other DNA clones. The recent report that the previously sequenced clone ARV-2 is also functional provides further support for these conclusions (55).

The identification of additional functional clones of HTLV-III/LAV provide clues for unravelling the biochemical basis of virus replication and cell killing. Clone HXB2gpt revealed attenuated cytopathic effects after transfection into cord blood mononuclear cells, which correlated with a delay in the appearance of viral particles. However, infection with virus derived from this plasmid clone revealed substantial cytopathic effects. Thus, the results in cord blood mononuclear cells are most likely due to either a lower transfection efficiency achievable with this clone, or mildly reduced infectivity of virus obtained from this clone in cord blood mononuclear cells, rather than a true reduction in the cytopathic potential of this genome per se.

Clone HXB3 also gives rise to cytopathic virus after transfection into cord blood mononuclear cells. The kinetics of virus production and cell killing are comparable to those of cultures transfected with HXB2gpt2. Virus derived from HXB3 also showed substantial cytopathic effects on ATH8 cells. Though only a portion of the sequence of HXB3 has been determined, it appears to be very similar to that of HXB2 (23,54). However, clone HXB2 has a termination codon at amino acid codon 124 in the 206 codon 3'orf gene, whereas clone HXB3 has a tryptophan codon at this position (54). The normal 3'orf product has been shown to be 27 kilodaltons (kd), whereas that generated from HXB2 is truncated and is 13 kd (42). This suggests that functions encoded by the second half of the 3'orf gene are nonessential. This is confirmed by the demonstration of similar functional activity of clone HXB2/3gpt in which only amino acids 43-206 of HXB2 3'orf are replaced by those derived from HXB3. Recent data have also demonstrated that deletions and frameshifts between amino acid codons 22 and 58 of 3'orf also do not affect the functional capabilities of the DNA clone (56). Thus, 3'orf is not required for in vitro replication or cytopathic activity.

The establishment of an in vitro system for AIDS, the identification of functional clones of HTLV-III/LAV, and the determination of the complete nucleotide sequence of functional clones provide the necessary tools and information for dissection of the viral genome to uncover the determinants of cytopathic activity and virus replication. By manipulating the genome of HXB2 to produce mutations in the virus, and analyzing the effects of such alterations in our in vitro system, we have recently mapped a major determinant of the cytopathic activity of HTLV-III/LAV to the 3' region of the virus (56). The use of molecular clones AIDS virus as shown here will also provide homogeneous stocks of virus useful for analysis of viral targets of cellular and humoral immunity.

#### ACKNOWLEDGEMENTS

This work was supported by a grant to L.R. from the American Foundation for AIDS Research.

The authors thank G. Shaw, B. Hahn, B. Starcich, S.F. Josephs, S. Broder, and J.B. Thielan for assistance and helpful discussions, M. Feinberg for the gift of HXB2gpt2, and M. Gonda for the electron micrographs.

# REFERENCES

1. CURRAN, J.W. (1985). *Ann. Intern. Med.* 103, 657.
2. CURRAN, J.W., MORGAN, M., HARDY, A.M., JAFFE, H.W., DARROW, W.W., and DOWDLE, W.R. (1985). *Science* 229, 1352.
3. LEVY, J.A., HOFFMAN, A.D., KRAMER, S.M., LANDIS, J.A., SHIMOBUKURO, J.M., and OSHIRO, L.S. (1984). *Science* 225, 840.
4. COFFIN, J., HAASE, A., LEVY, J.A., MONTAGNIER, L., OROSZOLAN, S., TEICH, N., TEMIN, H., TOYOSHIMA, K., VARMUS, H., VOGT, P., and WEISS, R. (1986). *Nature* 321, 10.
5. REDFIELD, R.R., WRIGHT, D.C., and TRAMONT, E.C. (1986). *N. Engl. J. Med.* 314, 131.
6. SIVAK, S.L., and WORMSER, G.P. (1985). *N. Engl. J. Med.* 313, 1352.
7. BARRÉ-SINOUSSE, F., CHERMANN, J.-C., REY, R., NUGEYRE, M.T., CHAMARET, S., GRUEST, J., DAUGUET, C., AXLER-BLIN, C., VÉZINET-BRUN, R., ROUZIOUX, C., ROZENBAUM, W., and MONTAGNIER, L. (1983). *Science* 220, 868.
8. CURRAN, J.W., LAWRENCE, D.N., JAFFE, H., KAPLAN, J.E., ZYLA, L.D., CHAMBERLAND, M., WEINSTEIN, R., LUI, K.-J., SCHONBERGER, L.B., SPIRA, T.J., ALEXANDER, J., SWINGER, G., AMANN, A., SOLOMON, S., AUERBACH, O., MILDVAN, D., STONEBURNER, R., JASON, J.M., HAVERKOS, H.W., and EVATT, B.L. (1984). *N. Engl. J. Med.* 310, 69.
9. FEORINO, P.M., KALYANARAMAN, V.S., HAVERKOS, H.W., CABRADILLA, C.D., WARFIELD, D.T., JAFFE, H.W., HARRISON, A.K., GOTTLIEB, M.S., GOLDINGER, D., CHERMANN, J.-C., BARRÉ-SINOUSSE, F., SPIRA, T., McDUGAL, J.S., CURRAN, J.W., MONTAGNIER, L., MURPHY, F.A., and FRANCIS, D.P. (1984). *Science* 225, 69-72.
10. GALLO, R.C., SALAHUDDIN, S.Z., POPOVIC, M., SHEARER, G.M., KAPLAN, M., HAYNES, B.F., PALKER, T.J., REDFIELD, R., OLESKE, J., SAFAI, B., WHITE, G., FOSTER, P., and MARKHAM, P.D. (1984). *Science* 224, 500.
11. KALYANARAMAN, V.S., CABRADILLA, C.D., GETCHELL, J.P., NARAYANAN, R., BRAFF, E.H., CHERMANN, J.-C., BARRÉ-SINOUSSE, F., MONTAGNIER, L., SPIRA, T.J., KAPLAN, J., FISHBEIN, D., JAFFE, H.W., CURRAN, J.W., and FRANCIS, D.P. (1984). *Science* 225, 321.
12. KAMINSKY, L.S., McHUGH, T., STITES, D., VOLDERBING, P., HENLE, G., HENLE, W., and LEVY, J.A. (1985). *Proc. Natl. Acad. Sci. USA* 82, 5535.
13. KITCHEN, L.W., BARIN, F., SULLIVAN, J.L., McLANE, M.F., BRETTLE, D.B., LEVINE, P.H., and ESSEX, M. (1984). *Nature* 312, 367.
14. LAURENCE, J., BRUN-VÉZINET, F., SCHUTZER, S.E., ROUZIOUX, C., KLATZMANN, D., BARRÉ-SINOUSSE, F., CHERMANN, J.-C., and MONTAGNIER, L. (1984). *N. Engl. J. Med.* 311, 1269.
15. SAFAI, B., GROOPMAN, J.E., POPOVIC, M., SCHUPBACH, J., SARNGADHARAN, M.G., ARNETT, K., SLISKI, A., and GALLO, R.C. (1984). *Lancet* 1, 1438.

16. S  
D  
P  
17. S  
(  
18. B  
7  
19. K  
G  
J  
20. P  
e  
21. A  
M  
22. B  
F  
9  
23. C  
S  
24. H  
R  
25. H  
S  
26. L  
N  
27. M  
ar  
28. R  
S  
I  
T  
N  
29. S  
M  
ar  
30. S  
H  
31. W  
Ce  
32. W  
MA  
33. S  
an

Bröder,  
for the

W., and  
, J.M.,

TEICH,  
(1986).

J. Med.

ET, S.,  
IX, C.,

CHAM-  
T.J.,  
LDVAN,  
(1984).

, WAR-  
R, D.,  
URRAN,  
cience

, M.,  
WHITE,

, R.,  
SPIRA,  
UNCIS,

ENLE,

D.B.,

MANN,  
N.

1.G.,

16. SALAHUDDIN, S.Z., MARKHAM, P.D., POPOVIC, M., SARNGADHARAN, M.G., ORN-DORFF, S., FLADAGAR, A., PATEL, A., GOLD, J., and GALLO, R.C. (1985). *Proc. Natl. Acad. Sci. USA* 82, 5530.
17. SARNGADHARAN, M.G., POPOVIC, M., BRUCH, L., SCHUPBACH, J., and GALLO, R.C. (1984). *Science* 224, 506.
18. BOWEN, D.L., LANE, H.C., and FAUCI, A.S. (1985). *Ann. Intern. Med.* 103, 704.
19. KLATZMANN, D., BARRÉ-SINOUSSE, F., NUGEYRE, M.T., DAUGUET, C., VILMER, E., GRISCELLI, C., BRUN-VÉZINET, F., ROUZIUX, C., GLUCKMAN, J.E., CHERMANN, J.-C., and MONTAGNIER, L. (1984). *Science* 225, 59.
20. POPOVIC, M., SARNGADHARAN, M.G., READ, E., and GALLO, R.C. (1984). *Science* 224, 497.
21. ALIZON, M., SONIGO, P., BARRÉ-SINOUSSE, F., CHERMANN, J.-C., TIOLLAIS, P., MONTAGNIER, L., and WAIN-HOBSON, S. (1984). *Nature* 312, 757.
22. BENN, S., RUTLEDGE, R., FOLKS, T., GOLD, J., BAKER, L., MCCORMICK, J., FEORINO, P., PIOT, P., QUINN, T., and MARTIN, M. (1985). *Science* 230, 949.
23. CROWL, R., GANGULY, K., GORDON, M., CONROY, R., SCHABER, M., KRAMER, R., SHAW, G., WONG-STAAAL, F., and REDDY, E.P. (1985). *Cell* 41, 979.
24. HAHN, B.H., GONDA, M.A., SHAW, G.M., POPOVIC, M., HOXIE, J.A., GALLO, R.C., and WONG-STAAAL, F. (1985). *Proc. Natl. Acad. Sci. USA* 82, 4813.
25. HAHN, B.H., SHAW, G.M., ARYA, S.K., POPOVIC, M., GALLO, R.C., and WONG-STAAAL, F. (1984). *Nature* 312, 166.
26. LUCIW, P.A., POTTER, S.J., STEIMER, K., DINA, D., and LEVY, J. (1984). *Nature* 312, 760.
27. MUESING, M.A., SMITH, D.A., CABRADILLA, C.D., BENTON, C.V., LASKY, L.A., and CAPON, D.J. (1985). *Nature* 313, 450.
28. RATNER, L., HASELTINE, W., PATARCA, R., LIVAK, K.J., STARCICH, B., JOSEPHS, S.F., DORAN, E.R., RAFALSKI, J.A., WHITEHORN, E.A., BAUMEISTER, K., IVANOFF, L., PETTEWAY, S.R., PEARSON, M.L., LAUTENBERGER, J.A., PAPAS, T.S., GHRAIEB, J., CHANG, N.T., GALLO, R.C., and WONG-STAAAL, F. (1985). *Nature* 313, 277-284.
29. SANCHEZ-PESCADOR, R., POWER, M.D., BARR, P.J., STEIMER, K.S., STEMPIEN, M.M., BROWN-SHIMER, S.L., GEE, W.W., RENARD, A., RANDOLPH, A., LEVY, J.A., and LUCIW, P.A. (1985). *Science* 227, 484.
30. STARCICH, B.R., HAHN, B.H., SHAW, G.M., MODROW, S., JOSEPHS, S.F., WOLF, H., GALLO, R.C., and WONG-STAAAL, F. (1986). *Cell* 45, 637.
31. WAIN-HOBSON, S., SONIGO, P., DANOS, O., COLE, S., and ALIZON, M. (1985). *Cell* 40, 9.
32. WONG-STAAAL, F., SHAW, G.M., HAHN, B.H., SALAHUDDIN, S.Z., POPOVIC, M., MARKHAM, P., REDFIELD, R., and GALLO, R.C. (1985). *Science* 229, 759.
33. SCHUPBACH, J., POPOVIC, M., GILDEN, R.V., GONDA, M.A., SARNGADHARAN, M.G., and GALLO, R.C. (1984). *Science* 224, 503.

34. SIEGEL, L.J., RATNER, L., JOSEPHS, S.F., DERSE, D., FEINBERG, M., REYES, G., O'BRIEN, S.J., and WONG-STAAAL, F. (1986). *Virology* 148, 226.
35. ARYA, S.K., GUO, C., JOSEPHS, S.F., and WONG-STAAAL, F. (1985). *Science* 229, 69.
36. DAYTON, A.I., SODROSKI, J.G., ROSEN, C.A., GOH, W.C., and HASELTINE, W.A. (1986). *Cell* 44, 941.
37. FISHER, A.G., FEINBERG, M.B., JOSEPHS, S.F., HARPER, M.E., MARSELLE, L.M., REYES, G., GONDA, M.A., ALDOVINI, A., DEBOUK, C., GALLO, R.C., and WONG-STAAAL, F. (1986). *Nature* 320, 367.
38. ROSEN, C.A., SODROSKI, J.G., GOH, W.C., DAYTON, A.I., LIPPKE, J., and HASELTINE, W.A. (1986). *Nature* 319, 555.
39. ROSEN, C.A., SODROSKI, J.G., and HASELTINE, W.A. (1985). *Cell* 41, 813.
40. SODROSKI, J.G., ROSEN, C.A., WONG-STAAAL, F., SALAHUDDIN, S.Z., POPOVIC, M., ARYA, S., GALLO, R.C., and HASELTINE, W.A. (1985). *Science* 225, 171.
41. SODROSKI, J., GOH, W.C., ROSEN, C., DAYTON, A., TERWILLIGER, E., and HASELTINE, W. (1986). *Nature* 321, 412.
42. FEINBERG, M.B., JARRETT, R.F., ALDOVINI, A., GALLO, R.C., and WONG-STAAAL, F. (1986). *Cell* 46, 807.
43. ALLAN, J., COLIGAN, J.E., LEE, T.H., McLANE, M.F., KANKI, P.J., GROOPMAN, J.E., and ESSEX, M. (1985). *Science* 230, 810.
44. KAN, N.C., FRANCHINI, G., WONG-STAAAL, F., DuBOIS, G.C., ROBEY, W.G., LAUTENBERGER, J.A., and PAPAS, T.C. (1986). *Science* 231, 1553.
45. LEE, T.-H., COLIGAN, J.E., ALLAN, J.S., McLANE, M.F., GROOPMAN, J.E., and ESSEX, M. (1986). *Science* 231, 1546.
46. FISHER, A.M., COLLALTI, E., RATNER, L., GALLO, R.C., and WONG-STAAAL, F. (1985). *Nature* 316, 262.
47. SHAW, G.M., HAHN, B.H., ARYA, S.K., GROOPMAN, J.E., GALLO, R.C., and WONG-STAAAL, F. (1984). *Science* 226, 1165.
48. MULLIGAN, R.C., and BERG, P. (1981). *Proc. Natl. Acad. Sci. USA* 78, 2072.
49. MAXAM, A.M., and GILBERT, W. (1980). *Methods Enzymol.* 65, 499.
50. SANGER, F., NICKLEN, S., and COULSON, A.R. (1977). *Proc. Natl. Acad. Sci. USA* 74, 5463.
51. MITSUYA, H., and BRODER, S. (1986). *Proc. Natl. Acad. Sci. USA* 83, 1911.
52. MITSUYA, H., WEINHOLD, K.J., FURMAN, P.A., ST. CLAIR, M.H., LEHRMANN, S.N., GALLO, R.C., BOLOGNESI, D., BARRY, D.W., and BRODER, S. (1985). *Proc. Natl. Acad. Sci. USA* 82, 7096.
53. PATTERSON, M.K. (1979). *Methods Enzymol.* 58, 141.
54. RATNER, L., STARCICH, B., JOSEPHS, S.F., HAHN, B.H., REDDY, E.P., LIVAK, K.J., PETTEWAY, S.R., PEARSON, M.L., HASELTINE, W.A., ARYA, S.K., and WONG-STAAAL, F. (1985). *Nucl. Acids Res.* 13, 8219.

REYES,  
Science  
J. W.A.  
L.M.,  
WONG-  
and  
13.  
POVIC,  
171.  
id HA-  
STAAL,  
OPMAN,  
LAU-  
, and  
, F.  
WONG-  
072.  
Sci.  
11.  
.N.,  
roc.  
AK,  
and

55. LEVY, J.A., CHENG-MAYER, C., DINA, D., and LUCIW, P.A. (1986). Science 232, 998.
56. FISHER, A.G., RATNER, L., MITSUYA, H., MARSELLE, L.M., HARPER, M.E., BRODER, S., GALLO, R.C., and WONG-STAAAL, F. (1986). Science 233, 655.

Address reprint requests to:  
Dr. Lee Ratner  
Division of Hematology and Oncology  
Box 8125, 660 S. Euclid  
Washington University  
St. Louis, MO 63110



# AIDS

*Mary Ann Liebert, Inc. publishers*